ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ШТУЧНОГО ІНТЕЛЕКТУ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Програмний застосунок для розпізнавання літописів на основі NLP та OCR»

на здобуття освітнього ступеня бакалавра
зі спеціальності 122 Комп'ютерні науки
освітньо-професійної програми Штучний інтелект

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Данило АРТЕМОВ
(підпис)

Виконав:
здобувач вищої освіти Данило АРТЕМОВ
група ШІД-42

Керівник: к.н.т. доцент Максим ФЕСЕНКО

Рецензент: _____
науковий ступінь, вчене звання *Ім'я, ПРІЗВИЩЕ*

Київ 2025

ДЕРЖАВНИЙ УНІВЕРСИТЕТ

ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут Інформаційних технологій

Кафедра Штучний інтелект

Ступінь вищої освіти бакалавр

Спеціальність 122 Комп'ютерні науки

Освітньо-професійна програма штучний інтелект

ЗАТВЕРДЖУЮ
Завідувач кафедрою Ольга ЗІНЧЕНКО
«____» _____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Артемов Данило Павлович

1. Тема кваліфікаційної роботи: Програмний застосунок для розпізнавання літописів на основі NLP та OCR
керівник кваліфікаційної роботи к.н.т. доцент Максим ФЕСЕНКО
затверджені наказом Державного університету інформаційно-комунікаційних технологій

від «24» лютого 2025 р.№56

2. Строк подання кваліфікаційної роботи «23» травня 2025 р.

3. Вихідні дані до кваліфікаційної роботи:науково-технічна література, дані інтернет-мережі,мова програмування Python, середовище розроблення IDE PyCharm

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

1. Провести аналіз існуючих систем розпізнавання старовинних текстів.
2. Обрати технології,інструменти для реалізації системи розпізнавання старовинних текстів.
3. Розробити алгоритм роботи застосунку з урахуванням особливостей старовинних текстів.
4. Розробка інтерфейсу застосунку. Підтвердити можливість його реалізації.

5. Перелік ілюстративного матеріалу: *презентація*

Дата видачі завдання «25» лютого 2025 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Отримання завдання	25.02-20.03.2025	Виконано
2	Аналіз роботи, підбір літератури	20.03-30.03.2025	Виконано
3	Аналіз моделі “Tesseract”	30.03-09.04.2025	Виконано
4	Розробка методів моделі “Tesseract OCR”	09.04-19.04.2025	Виконано
5	Програмна реалізація	19.04-28.04.2025	Виконано
6	Оформлення пояснювальної записки	28.04-30.04.2025	Виконано
7	Перевірка на плагіат	30.04-06.05.2025	Виконано
8	Рецензування	06.05-14.05.2025	Виконано
9	Підготовка презентації	14.05-20.05.2025	Виконано
10	Попередній захист кваліфікаційної роботи	20.05-23.05.2025	Виконано

Здобувач вищої освіти _____ Данило АРТЕМОВ
(підпис)

Керівник
кваліфікаційної роботи _____ Максти ФЕСЕНКО
(підпис)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня бакалавра: 55 стор., 22 рис., 2 табл., 17 джерел.

Мета роботи - дослідити та розробити програмний застосунок для автоматизованого розпізнавання та аналізу історичних літописів за допомогою сучасних технологій оптичного розпізнавання символів (OCR) та обробки природної мови (NLP). Розроблений застосунок дозволяє швидко та точно обробляти великі обсяги історичних текстів, сприяючи збереженню культурної спадщини та полегшуючи доступ до історичних документів.

Об'єкт дослідження - процеси розпізнавання та обробки історичних літописів з використанням сучасних технологій штучного інтелекту.

Предмет дослідження - методи та алгоритми розпізнавання тексту (OCR), інструменти NLP для нормалізації та перекладу старовинних текстів, а також програмна реалізація системи обробки літописів.

Короткий зміст роботи: Робота присвячена створенню системи, що автоматизує процес розпізнавання та аналізу старовинних літописів. У першому розділі розглянуто теоретичні аспекти технологій OCR і NLP, а також особливості старовинних документів, таких як багатоколонний формат та архаїчні шрифти (Fraktur). Другий розділ описує реалізацію системи: вибір моделей та алгоритмів для OCR і NLP, архітектуру програмного застосунку та методи підготовки даних. Третій розділ присвячений практичній реалізації - тестуванню системи на прикладах історичних літописів, аналізу результатів розпізнавання та оцінці точності. У роботі особлива увага приділена адаптації моделей до складних особливостей літописів, а також перспективам подальшого розвитку системи.

КЛЮЧОВІ СЛОВА: OCR, NLP, ІСТОРИЧНІ ЛІТОПИСИ, ЦИФРОВА ОБРОБКА ТЕКСТІВ, НЕЙРОННІ МЕРЕЖІ, РОЗПІЗНАВАННЯ ТЕКСТУ, МАШИННЕ НАВЧАННЯ.

ABSTRACT

Text section of the qualification work for obtaining the bachelor's degree: 55 pages, 22 figures, 2 tables, 17 sources.

The purpose of the work is to research and develop a software application for automated recognition and analysis of historical chronicles using modern technologies of optical character recognition (OCR) and natural language processing (NLP). The developed application allows you to quickly and accurately process large volumes of historical texts, contributing to the preservation of cultural heritage and facilitating access to historical documents.

The object of the research is the processes of recognition and processing of historical chronicles using modern technologies of artificial intelligence.

The subject of the research is methods and algorithms of text recognition (OCR), NLP tools for normalization and translation of ancient texts, as well as software implementation of the chronicle processing system.

Summary of the work: The work is devoted to the creation of a system that automates the process of recognition and analysis of ancient chronicles. The first section discusses the theoretical aspects of OCR and NLP technologies, as well as the features of ancient documents, such as multi-column format and archaic fonts (Fraktur). The second section describes the implementation of the system: the choice of models and algorithms for OCR and NLP, the architecture of the software application and data preparation methods. The third section is devoted to practical implementation - testing the system on examples of historical chronicles, analyzing the recognition results and assessing the accuracy. The paper pays special attention to the adaptation of models to the complex features of chronicles, as well as the prospects for further development of the system.

KEYWORDS: OCR, NLP, HISTORICAL CHRONICLES, DIGITAL TEXT PROCESSING, NEURAL NETWORKS, TEXT RECOGNITION, MACHINE LEARNING.

ЗМІСТ

ВСТУП

Актуальність теми. У сучасному світі цифрова трансформація історичних документів є важливим напрямком збереження культурної спадщини та полегшення доступу до історичних даних. Літописи, як ключові джерела інформації про минуле, часто зберігаються у вигляді рукописів або старих друкованих текстів, що ускладнює їх аналіз та використання дослідниками, істориками та широкою громадськістю.

Мета дослідження. Метою дослідження є розробка програмного застосунку для автоматизованого розпізнавання та аналізу текстів літописів з використанням технологій оптичного розпізнавання символів (OCR) та обробки природної мови (NLP). Додаток має на меті забезпечити ефективне оцифрування історичних документів, підвищити точність розпізнавання тексту та автоматизувати їх семантичний аналіз для полегшення доступу до культурної спадщини та підтримки історичних та лінгвістичних досліджень.

Об’єкт дослідження. Об'єктом дослідження є історичні літописи, представлені у вигляді рукописів або стародрукованих текстів, а також процеси їх автоматизованої оцифровки, розпізнавання та семантичного аналізу за допомогою технологій оптичного розпізнавання символів (OCR) та обробки природної мови (NLP).

Предмет дослідження. Предметом дослідження є методи, алгоритми та програмні засоби, що забезпечують автоматизоване розпізнавання текстів історичних літописів за допомогою технологій оптичного розпізнавання символів (OCR) та обробки природної мови (NLP), та їх інтеграція у функціональні програмні додатки для оцифрування, аналізу та збереження вмісту літопису.

Методика дослідження. Для досягнення мети дослідження застосовується комплексна методика, що поєднує теоретичний аналіз літератури та існуючих рішень для вибору оптимальних технологій OCR і NLP, збір і підготовку даних (оцифрування літописів, обробка зображень, анотація текстів), розробку та навчання моделей OCR і NLP з використанням машинного та глибокого навчання, проектування модульної архітектури застосунку з інтуїтивним інтерфейсом,

тестування системи (модульне, інтеграційне, користувацьке) для оцінки точності та зручності, оптимізацію алгоритмів і продуктивності, а також документування результатів і апробацію в реальних умовах (архівах, бібліотеках), що забезпечує створення ефективного інструменту для оцифровки, розпізнавання та аналізу історичних текстів із практичною цінністю для цифрових гуманітарних наук.

Наукова новизна. Наукова новизна дослідження полягає у розробці інтегрованого програмного рішення, яке поєднує адаптовані технології оптичного розпізнавання символів та обробки природної мови для автоматизованої обробки історичних літописів.

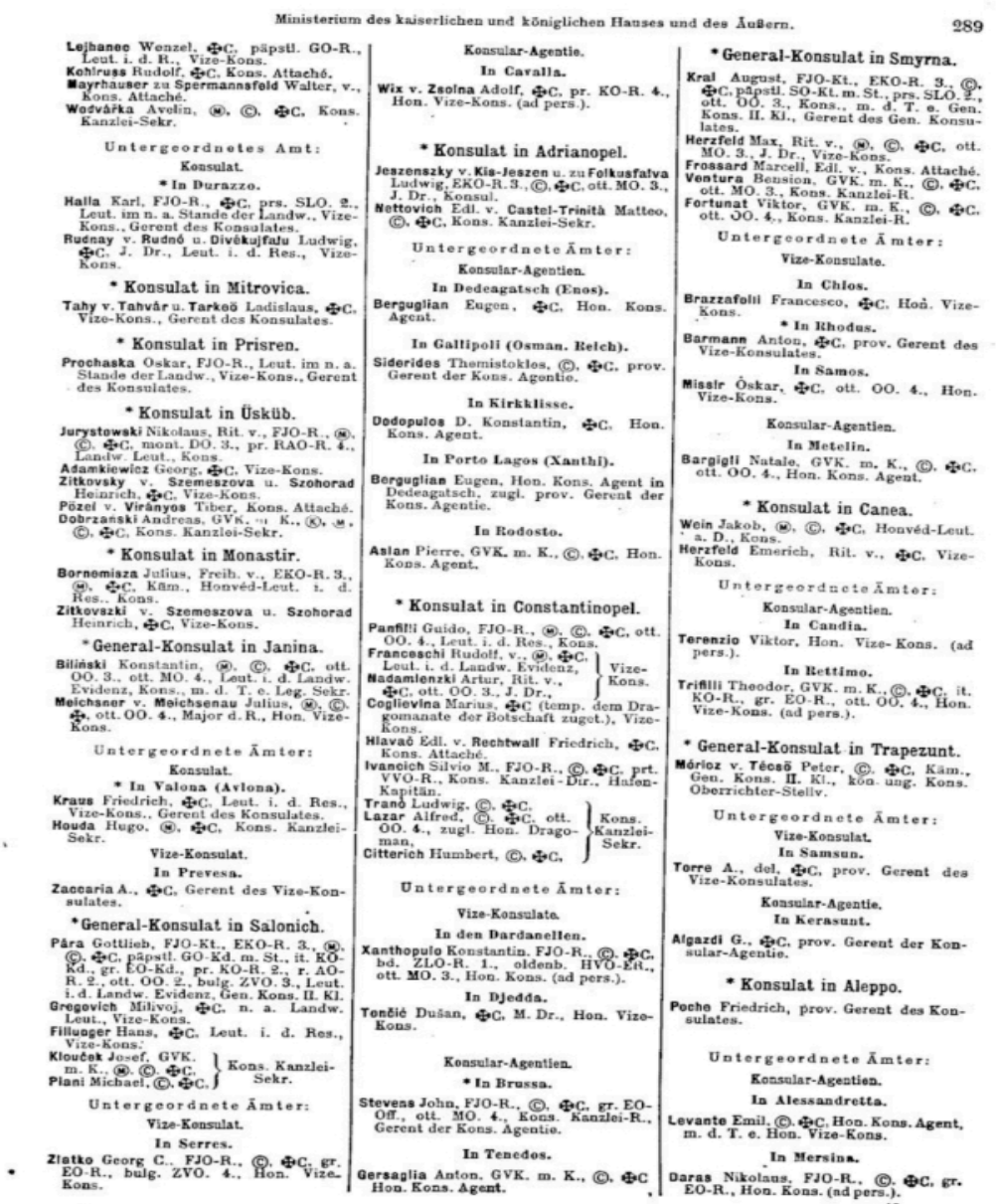
Практична значущість результатів роботи. Результати дослідження не лише вирішують практичні завдання оцифрування та аналізу літописів, але й створюють платформу для подальшого розвитку цифрових технологій у гуманітарних науках. Впровадження програмного застосунку в архівах, бібліотеках і дослідницьких установах сприятиме збереженню та популяризації культурної спадщини, підвищенню ефективності наукових досліджень і формуванню нових міждисциплінарних підходів до вивчення історичних джерел.

1 ТЕОРЕТИЧНІ ОСНОВИ РОЗПІЗНАВАННЯ ТА ОБРОБКИ ЛІТОПИСНИХ ТЕКСТІВ

1.1 Актуальність розпізнавання історичних текстів

Обробка історичних документів, таких як багатотомна серія "Hof- und Staatsschematismus" Габсбурзької монархії (1702-1918), що включає 145 томів і близько 150,000 сторінок, є дуже складним завданням через специфічні особливості цих джерел. У цих документах зафіксовані детальні відомості про адміністративну ієрархію, службові кар'єри та соціальні зв'язки еліт Центральної Європи того часу. Проте, через багатоколонний формат, складні ієрархічні списки, наявність спеціальних символів та використання старовинних шрифтів, автоматизоване вилучення інформації з них стає особливо важким. Стандартні системи оптичного розпізнавання тексту (OCR), наприклад, Tesseract, не справляються з такими складними матеріалами - вони дають низьку якість результатів. Водночас ручне переписування тексту з таких документів потребує дуже багато часу та часто супроводжується помилками.

Дослідження Флейшхакера, Керна та Гьодерле (2025)[4,5] показує необхідність сучасних підходів, зокрема використання машинного навчання, для обробки подібних документів. Їхній аналіз на прикладі "Schematismus" показує, наскільки складною є структура цих текстів: багатоколонне розташування, ієрархічні зв'язки та спеціальні символи (наприклад, фігурні дужки), які потребують нових рішень у сфері цифрової обробки (рис. 1.1).



Ministerium des kaiserlichen und königlichen Hauses und des Äußern. 289

Lejhanec Wenzel, ✠C. päpstl. GO-R., Leut. i. d. R., Vize-Kons.
Kohlruss Rudolf, ✠C. Kons. Attaché.
Mayrhauser zu Spermannsfeld Walter, v., Kons. Attaché.
Wodvářka Avelin, Ⓜ, Ⓒ, ✠C. Kons. Kanzlei-Sekr.

Untergeordnetes Amt:

Konsulat.

* In Durazzo.

Halla Karl, FJO-R., ✠C. prs. SLO. 2., Leut. im n. a. Stande der Landw., Vize-Kons., Gerent des Konsulates.
Rudnay v. Rudnó u. Divékujfalu Ludwig, ✠C. J. Dr., Leut. i. d. Res., Vize-Kons.

* Konsulat in Mitrovica.

Tahy v. Tahvár u. Tarkeö Ladislaus, ✠C. Vize-Kons., Gerent des Konsulates.

* Konsulat in Prisren.

Prochaska Oskar, FJO-R., Leut. im n. a. Stande der Landw., Vize-Kons., Gerent des Konsulates.

* Konsulat in Üsküb.

Jurystowski Nikolaus, Rit. v., FJO-R., Ⓜ, Ⓒ, ✠C. mont. DO. 3., pr. RAO-R. 4., Landw. Leut., Kons.
Adamkiewicz Georg, ✠C. Vize-Kons.
Zitkovsky v. Szemeszova u. Szohorad Heinrich, ✠C. Vize-Kons.
Pözel v. Virányos Tiber, Kons. Attaché.
Dobrzański Andreas, GVK. m. K., Ⓚ, Ⓜ, Ⓒ, ✠C, Kons. Kanzlei-Sekr.

* Konsulat in Monastir.

Bornemisza Julius, Freih. v., EKO-R. 3., Ⓜ, ✠C. Käm., Honvéd-Leut. i. d. Res., Kons.
Zitkovszki v. Szemeszova u. Szohorad Heinrich, ✠C. Vize-Kons.

* General-Konsulat in Janina.

Biliński Konstantin, Ⓜ, Ⓒ, ✠C. ott. OO. 3., ott. MO. 4., Leut. i. d. Landw. Evidenz, Kons., m. d. T. e. Leg. Sekr.
Meichsner v. Meichsenau Julius, Ⓜ, Ⓒ, ✠, ott. OO. 4., Major d. R., Hon. Vize-Kons.

Untergeordnete Ämter:

Konsulat.

* In Valona (Avlona).

Kraus Friedrich, ✠C. Leut. i. d. Res., Vize-Kons., Gerent des Konsulates.
Houda Hugo, Ⓜ, ✠C. Kons. Kanzlei-Sekr.

Vize-Konsulat.

In Prevesa.

Zaccaria A., ✠C. Gerent des Vize-Konsulates.

* General-Konsulat in Salonich.

Pára Gottlieb, FJO-Kt., EKO-R. 3., Ⓜ, Ⓒ, ✠C. päpstl. GO-Kd. m. St., it. KO-Kd., gr. EO-Kd., pr. KO-R. 2., r. AO-R. 2., ott. OO. 2., bulg. ZVO. 3., Leut. i. d. Landw. Evidenz, Gen. Kons. II. Kl.
Gregovich Milivoj, ✠C. n. a. Landw. Leut., Vize-Kons.
Fillunger Hans, ✠C. Leut. i. d. Res., Vize-Kons.
Klouček Josef, GVK. m. K., Ⓜ, Ⓒ, ✠C. } Kons. Kanzlei-Sekr.
Pisni Michael, Ⓒ, ✠C. }

Untergeordnete Ämter:

Vize-Konsulat.

In Serres.

Zlatko Georg C., FJO-R., Ⓒ, ✠C. gr. EO-R., bulg. ZVO. 4., Hon. Vize-Kons.

Konsular-Agentie.

In Cavalla.

Wix v. Zsolna Adolf, ✠C. pr. KO-R. 4., Hon. Vize-Kons. (ad pers.).

* Konsulat in Adrianopel.

Jeszenszky v. Kis-Jeszen u. zu Folkusfalva Ludwig, EKO-R. 3., Ⓒ, ✠C. ott. MO. 3., J. Dr., Konsul.
Nottovich Edl. v. Castel-Trinità Matteo, Ⓒ, ✠C. Kons. Kanzlei-Sekr.

Untergeordnete Ämter:

Konsular-Agentien.

In Dedeagatsch (Enos).

Berguglian Eugen, ✠C. Hon. Kons. Agent.

In Gallipoli (Osman. Reich).

Siderides Themistokles, Ⓒ, ✠C. prov. Gerent der Kons. Agentie.

In Kirkklisse.

Dodopulos D. Konstantin, ✠C. Hon. Kons. Agent.

In Porto Lagos (Xanthi).

Berguglian Eugen, Hon. Kons. Agent in Dedeagatsch, zugl. prov. Gerent der Kons. Agentie.

In Rodosto.

Aslan Pierre, GVK. m. K., Ⓒ, ✠C. Hon. Kons. Agent.

* Konsulat in Constantinopel.

Panfilli Guido, FJO-R., Ⓜ, Ⓒ, ✠C, ott. OO. 4., Leut. i. d. Res., Kons.
Franceschi Rudolf, v., Ⓜ, ✠C. Leut. i. d. Landw. Evidenz, } Vize-
Nadamlenzki Artur, Rit. v., } Kons.
✠C, ott. OO. 3., J. Dr.,
Coglievina Marius, ✠C (temp. dem Dragomanate der Botschaft zuget.), Vize-Kons.
Hlavač Edl. v. Rechtwall Friedrich, ✠C. Kons. Attaché.
Ivancich Silvio M., FJO-R., Ⓒ, ✠C. prt. VVO-R., Kons. Kanzlei-Dir., Hafen-Kapitän.
Tranč Ludwig, Ⓒ, ✠C. } Kons.
Lazar Alfred, Ⓒ, ✠C. ott. } Kanzlei-
OO. 4., zugl. Hon. Dragoman, } Sekr.
Citterich Humbert, Ⓒ, ✠C. }

Untergeordnete Ämter:

Vize-Konsulate.

In den Dardanellen.

Xanthopulo Konstantin, FJO-R., Ⓒ, ✠C. bd. ZLO-R. 1., oldenb. HVO-ER., ott. MO. 3., Hon. Kons. (ad pers.).

In Djedda.

Tončić Dušan, ✠C. M. Dr., Hon. Vize-Kons.

Konsular-Agentien.

* In Brussa.

Stevens John, FJO-R., Ⓒ, ✠C. gr. EO-Off., ott. MO. 4., Kons. Kanzlei-R., Gerent der Kons. Agentie.

In Tenedos.

Gersaglia Anton, GVK. m. K., Ⓒ, ✠C Hon. Kons. Agent.

* General-Konsulat in Smyrna.

Kral August, FJO-Kt., EKO-R. 3., Ⓒ, ✠C. päpstl. SO-Kt. m. St., prs. SLO. 2., ott. OO. 3., Kons., m. d. T. e. Gen. Kons. II. Kl., Gerent des Gen. Konsulates.
Herzfeld Max, Rit. v., Ⓜ, Ⓒ, ✠C. ott. MO. 3., J. Dr., Vize-Kons.
Frossard Marcell, Edl. v., Kons. Attaché.
Ventura Benion, GVK. m. K., Ⓒ, ✠C. ott. MO. 3., Kons. Kanzlei-R.
Fortunat Viktor, GVK. m. K., Ⓒ, ✠C. ott. OO. 4., Kons. Kanzlei-R.

Untergeordnete Ämter:

Vize-Konsulate.

In Chios.

Brazzafolli Francesco, ✠C. Hon. Vize-Kons.

* In Rhodus.

Barmann Anton, ✠C. prov. Gerent des Vize-Konsulates.

In Samos.

Missir Oskar, ✠C. ott. OO. 4., Hon. Vize-Kons.

Konsular-Agentien.

In Metelin.

Bargigli Natale, GVK. m. K., Ⓒ, ✠C. ott. OO. 4., Hon. Kons. Agent.

* Konsulat in Canea.

Wein Jakob, Ⓜ, Ⓒ, ✠C. Honvéd-Leut. a. D., Kons.
Herzfeld Emerich, Rit. v., ✠C. Vize-Kons.

Untergeordnete Ämter:

Konsular-Agentien.

In Candia.

Terenzio Viktor, Hon. Vize-Kons. (ad pers.).

In Rettimo.

Triffili Theodor, GVK. m. K., Ⓒ, ✠C. it. KO-R., gr. EO-R., ott. OO. 4., Hon. Vize-Kons. (ad pers.).

* General-Konsulat in Trapezunt.

Mórioz v. Técsö Peter, Ⓒ, ✠C. Käm., Gen. Kons. II. Kl., kön. ung. Kons. Oberrichter-Stellv.

Untergeordnete Ämter:

Vize-Konsulat.

In Samsun.

Torre A., del, ✠C. prov. Gerent des Vize-Konsulates.

Konsular-Agentie.

In Kerasunt.

Algazdi G., ✠C. prov. Gerent der Konsular-Agentie.

* Konsulat in Aleppo.

Poche Friedrich, prov. Gerent des Konsulates.

Untergeordnete Ämter:

Konsular-Agentien.

In Alessandretta.

Levante Emil, Ⓒ, ✠C. Hon. Kons. Agent, m. d. T. e. Hon. Vize-Kons.

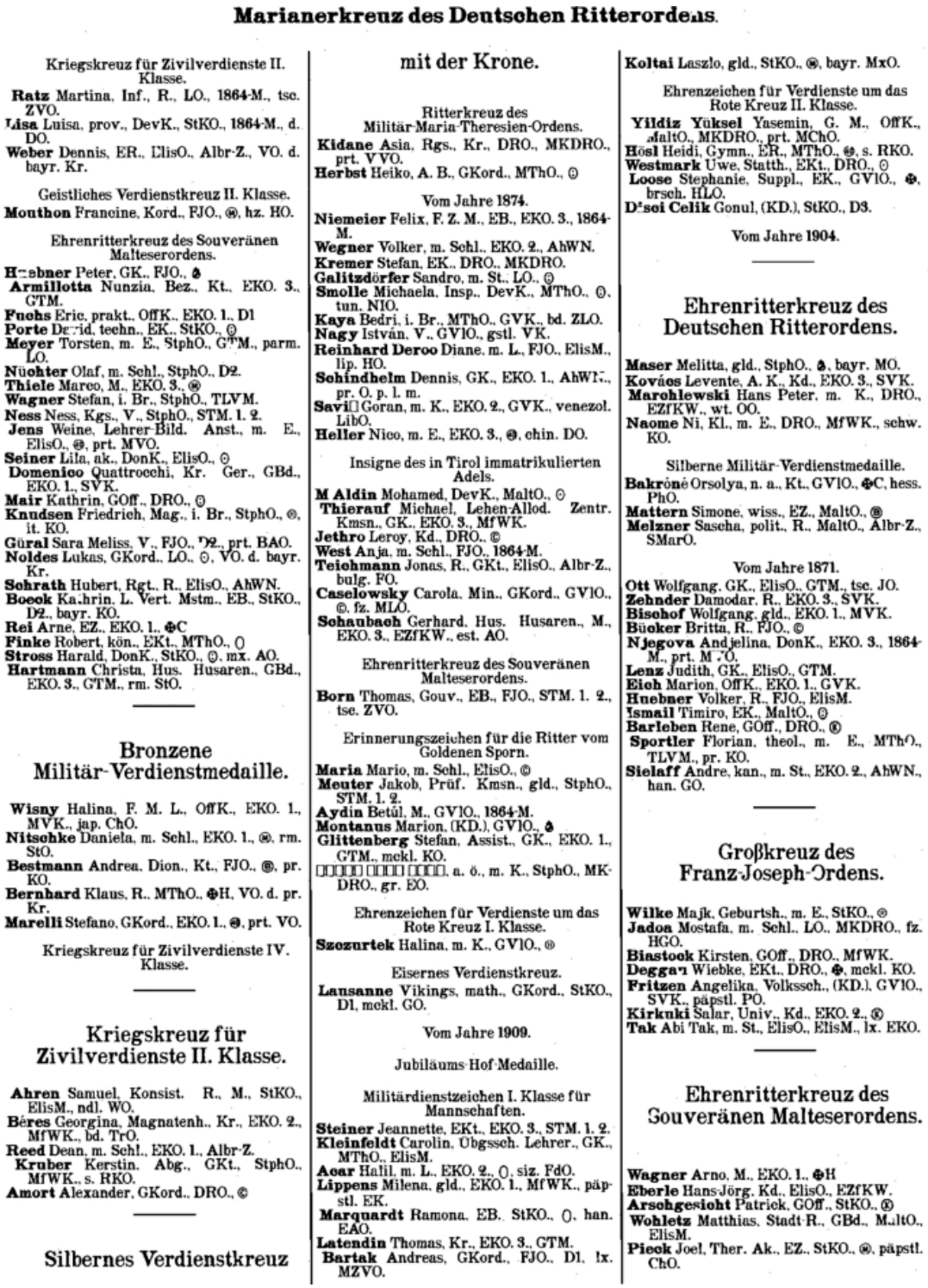
In Mersina.

Daras Nikolaus, FJO-R., Ⓒ, ✠C. gr. EO-R., Hon. Kons. (ad pers.).

19

Рис.1.1 - Приклад сторінки “Hof- und Staatsschematismus” 1910 року.

Автори пропонують інноваційний метод, що поєднує сегментацію сторінок на елементи макету за допомогою моделі Faster R-CNN і тонке налаштування OCR-двигуна Tesseract на кастомізований шрифт, адаптований до оригінального стилю документів. Для тренування моделі використано синтетичні дані, створені Python-скриптом із LaTeX-кодом, які імітують символи, макет та шрифти “Schematismus”. Процес створення таких даних включає генерацію документів із різноманітними макетами (рис. 1.2).



Marianerkreuz des Deutschen Ritterordens.

Kriegskreuz für Zivilverdienste II. Klasse.

Ratz Martina, Inf., R., LO., 1864-M., tsc. ZVO.
Lisa Luisa, prov., DevK., StKO., 1864-M., d. DO.
Weber Dennis, ER., ElisO., Albr-Z., VO. d. bayr. Kr.

Geistliches Verdienstkreuz II. Klasse.

Mouthon Francine, Kord., FJO., ⊛, hz. HO.

Ehrenritterkreuz des Souveränen Malteserordens.

Hübner Peter, GK., FJO., ♁
Armillotta Nunzia, Bez., Kt., EKO. 3., GTM.
Fuchs Eric, prakt., OffK., EKO. 1., Dl
Porte David, techn., EK., StKO., ⊙
Meyer Torsten, m. E., StphO., GTM., parm. LO.
Nüchter Olaf, m. Schl., StphO., D2.
Thiele Marco, M., EKO. 3., ⊛
Wagner Stefan, i. Br., StphO., TLVM.
Ness Ness, Kgs., V., StphO., STM. 1. 2.
Jens Weine, Lehrer-Bild. Anst., m. E., ElisO., ⊛, prt. MVO.
Seiner Lila, ak., DonK., ElisO., ⊙
Domenico Quattrocchi, Kr. Ger., GBd., EKO. 1., SVK.
Mair Kathrin, GOff., DRO., ⊙
Knudsen Friedrich, Mag., i. Br., StphO., ⊛, it. KO.
Güral Sara Meliss, V., FJO., D2., prt. BAO.
Noldes Lukas, GKord., LO., ⊙, VO. d. bayr. Kr.
Schrath Hubert, Rgt., R., ElisO., AhWN.
Boeck Katrin, L. Vert. Mstm., EB., StKO., D2., bayr. KO.
Rei Arne, EZ., EKO. 1., ✠C
Finke Robert, kön., EKt., MThO., ⬡
Stross Harald, DonK., StKO., ⊙, mx. AO.
Hartmann Christa, Hus. Husaren., GBd., EKO. 3., GTM., rm. StO.

Bronzene Militär-Verdienstmedaille.

Wisny Halina, F. M. L., OffK., EKO. 1., MVK., jap. ChO.
Nitschke Daniela, m. Schl., EKO. 1., ⊛, rm. StO.
Bestmann Andrea, Dion., Kt., FJO., ⊛, pr. KO.
Bernhard Klaus, R., MThO., ✠H, VO. d. pr. Kr.
Marelli Stefano, GKord., EKO. 1., ⊛, prt. VO.

Kriegskreuz für Zivilverdienste IV. Klasse.

Kriegskreuz für Zivilverdienste II. Klasse.

Ahren Samuel, Konsist. R., M., StKO., ElisM., ndl. WO.
Béres Georgina, Magnatenh., Kr., EKO. 2., MfWK., bd. TrO.
Reed Dean, m. Schl., EKO. 1., Albr-Z.
Kruber Kerstin, Abg., GKt., StphO., MfWK., s. RKO.
Amort Alexander, GKord., DRO., ©

Silbernes Verdienstkreuz mit der Krone.

Ritterkreuz des Militär-Maria-Theresien-Ordens.

Kidane Asia, Rgs., Kr., DRO., MKDRO., prt. VVO.
Herbst Heiko, A. B., GKord., MThO., ⊙

Vom Jahre 1874.

Niemeier Felix, F. Z. M., EB., EKO. 3., 1864-M.
Wegner Volker, m. Schl., EKO. 2., AhWN.
Kremer Stefan, EK., DRO., MKDRO.
Galitzdörfer Sandro, m. St., LO., ⊙
Smolle Michaela, Insp., DevK., MThO., ⊙, tun. NIO.
Kaya Bedri, i. Br., MThO., GVK., bd. ZLO.
Nagy István, V., GVlO., gstl. VK.
Reinhard Deroo Diane, m. L., FJO., ElisM., lip. HO.
Schindhelm Dennis, GK., EKO. 1., AhWN., pr. O. p. l. m.
Savić Goran, m. K., EKO. 2., GVK., venezol. LibO.
Heller Nico, m. E., EKO. 3., ⊛, chin. DO.

Insigne des in Tirol immatrikulierten Adels.

M Aldin Mohamed, DevK., MaltO., ⊙
Thierauf Michael, Lehen-Allod. Zentr. Kmsn., GK., EKO. 3., MfWK.
Jethro Leroy, Kd., DRO., ©
West Anja, m. Schl., FJO., 1864-M.
Teichmann Jonas, R., GKt., ElisO., Albr-Z., bulg. FO.
Caselowsky Carola, Min., GKord., GVlO., ©, fz. MLO.
Schaubach Gerhard, Hus. Husaren., M., EKO. 3., EZfKW., est. AO.

Ehrenritterkreuz des Souveränen Malteserordens.

Born Thomas, Gouv., EB., FJO., STM. 1. 2., tsc. ZVO.

Erinnerungszeichen für die Ritter vom Goldenen Sporn.

Maria Mario, m. Schl., ElisO., ©
Meuter Jakob, Prüf. Kmsn., gld., StphO., STM. 1. 2.
Aydin Betül, M., GVlO., 1864-M.
Montanus Marion, (KD.), GVlO., ♁
Glittenberg Stefan, Assist., GK., EKO. 1., GTM., mekl. KO.
□□□□□ □□□□ □□□□, a. ö., m. K., StphO., MK-DRO., gr. EO.

Ehrenzeichen für Verdienste um das Rote Kreuz I. Klasse.

Szczurtek Halina, m. K., GVlO., ⊛

Eisernes Verdienstkreuz.

Lausanne Vikings, math., GKord., StKO., Dl, mekl. GO.

Vom Jahre 1909.

Jubiläums-Hof-Medaille.

Militärdienstzeichen I. Klasse für Mannschaften.

Steiner Jeannette, EKt., EKO. 3., STM. 1. 2.
Kleinfeldt Carolin, Übgsseh. Lehrer., GK., MThO., ElisM.
Acar Halil, m. L., EKO. 2., ⬡, siz. FdO.
Lippens Milena, gld., EKO. 1., MfWK., päpstl. EK.
Marquardt Ramona, EB., StKO., ⬡, han. EAO.
Latendin Thomas, Kr., EKO. 3., GTM.
Bartak Andreas, GKord., FJO., Dl, lx. MZVO.
Koltai Laszlo, gld., StKO., ⊛, bayr. MxO.

Ehrenzeichen für Verdienste um das Rote Kreuz II. Klasse.

Yildiz Yüksel Yasemin, G. M., OffK., MaltO., MKDRO., prt. MChO.
Hösl Heidi, Gymn., ER., MThO., ⊛, s. RKO.
Westmark Uwe, Statth., EKt., DRO., ⊙
Loose Stephanie, Suppl., EK., GVlO., ✠, brsch. HLO.
Dösei Celik Gonul, (KD.), StKO., D3.

Vom Jahre 1904.

Ehrenritterkreuz des Deutschen Ritterordens.

Maser Melitta, gld., StphO., ♁, bayr. MO.
Kovács Levente, A. K., Kd., EKO. 3., SVK.
Marohlewski Hans Peter, m. K., DRO., EZfKW., wt. OO.
Naome Ni, Kl., m. E., DRO., MfWK., schw. KO.

Silberne Militär-Verdienstmedaille.

Bakróné Orsolya, n. a., Kt., GVlO., ✠C, hess. PhO.
Mattern Simone, wiss., EZ., MaltO., ⊛
Melzner Sascha, polit., R., MaltO., Albr-Z., SMarO.

Vom Jahre 1871.

Ott Wolfgang, GK., ElisO., GTM., tsc. JO.
Zehnder Damodar, R., EKO. 3., SVK.
Bischof Wolfgang, gld., EKO. 1., MVK.
Bücker Britta, R., FJO., ©
Njegova Andjelina, DonK., EKO. 3., 1864-M., prt. MVO.
Lenz Judith, GK., ElisO., GTM.
Eich Marion, OffK., EKO. 1., GVK.
Huebner Volker, R., FJO., ElisM.
Ismail Timiro, EK., MaltO., ⊙
Barleben Rene, GOff., DRO., ⊛
Sportler Florian, theol., m. E., MThO., TLVM., pr. KO.
Sielaff Andre, kan., m. St., EKO. 2., AhWN., han. GO.

Großkreuz des Franz-Joseph-Ordens.

Wilke Majk, Geburtsh., m. E., StKO., ⊛
Jadoa Mostafa, m. Schl., LO., MKDRO., fz. HGO.
Biastock Kirsten, GOff., DRO., MfWK.
Deggau Wiebke, EKt., DRO., ✠, mekl. KO.
Fritzen Angelika, Volkssch., (KD.), GVlO., SVK., päpstl. PO.
Kirknki Salar, Univ., Kd., EKO. 2., ⊛
Tak Abi Tak, m. St., ElisO., ElisM., lx. EKO.

Ehrenritterkreuz des Souveränen Malteserordens.

Wagner Arno, M., EKO. 1., ✠H
Eberle Hans-Jörg, Kd., ElisO., EZfKW.
Arsehgesicht Patrick, GOff., StKO., ⊛
Wohletz Matthias, Stadt-R., GBd., MaltO., ElisM.
Pieck Joel, Ther. Ak., EZ., StKO., ⊛, päpstl. ChO.

Рис.1.2 - Приклад синтетичного документа в стилі “Schematismus”, використаного для тренування моделі Faster R-CNN.

Запропонований підхід забезпечує значне покращення якості OCR. Загалом, порівняно зі стандартним Tesseract, помилки знижено на 15.68% (CER) і 19.95% (WER) за допомогою тонкого налаштування та комбінації з сегментованим макетом. Процес сегментації, проілюстрований у схемі обробки елементів макету

(рис.1.3), дозволяє зберігати контекст тексту, автоматично сортувати блоки та вибірково вилучати елементи, такі як заголовки чи параграфи.

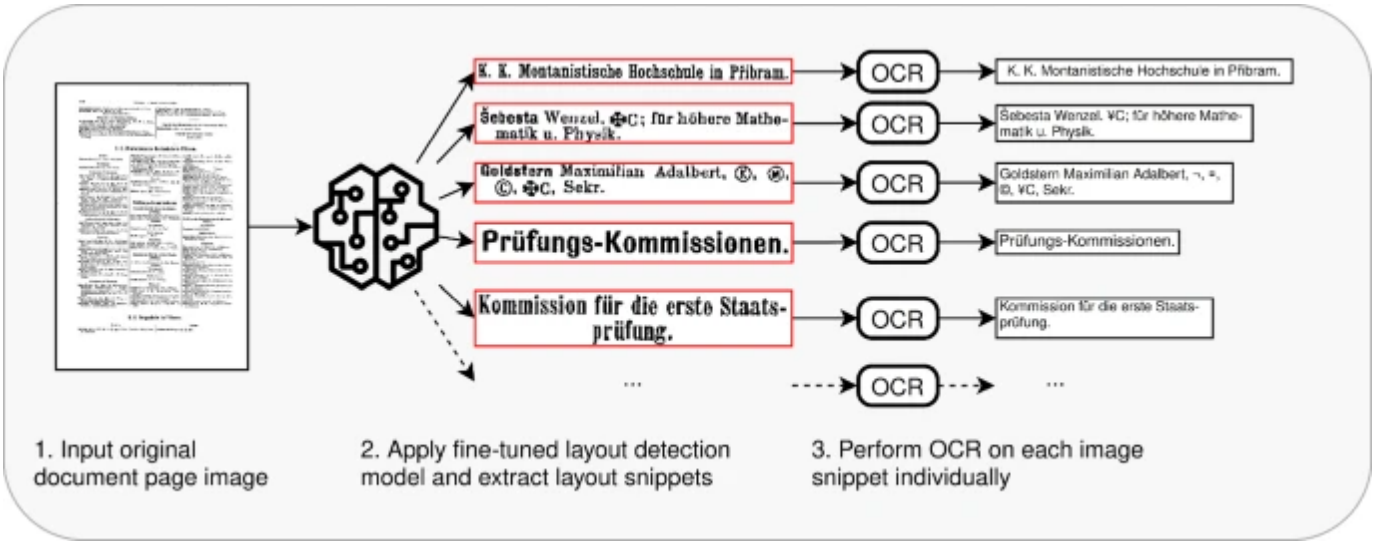
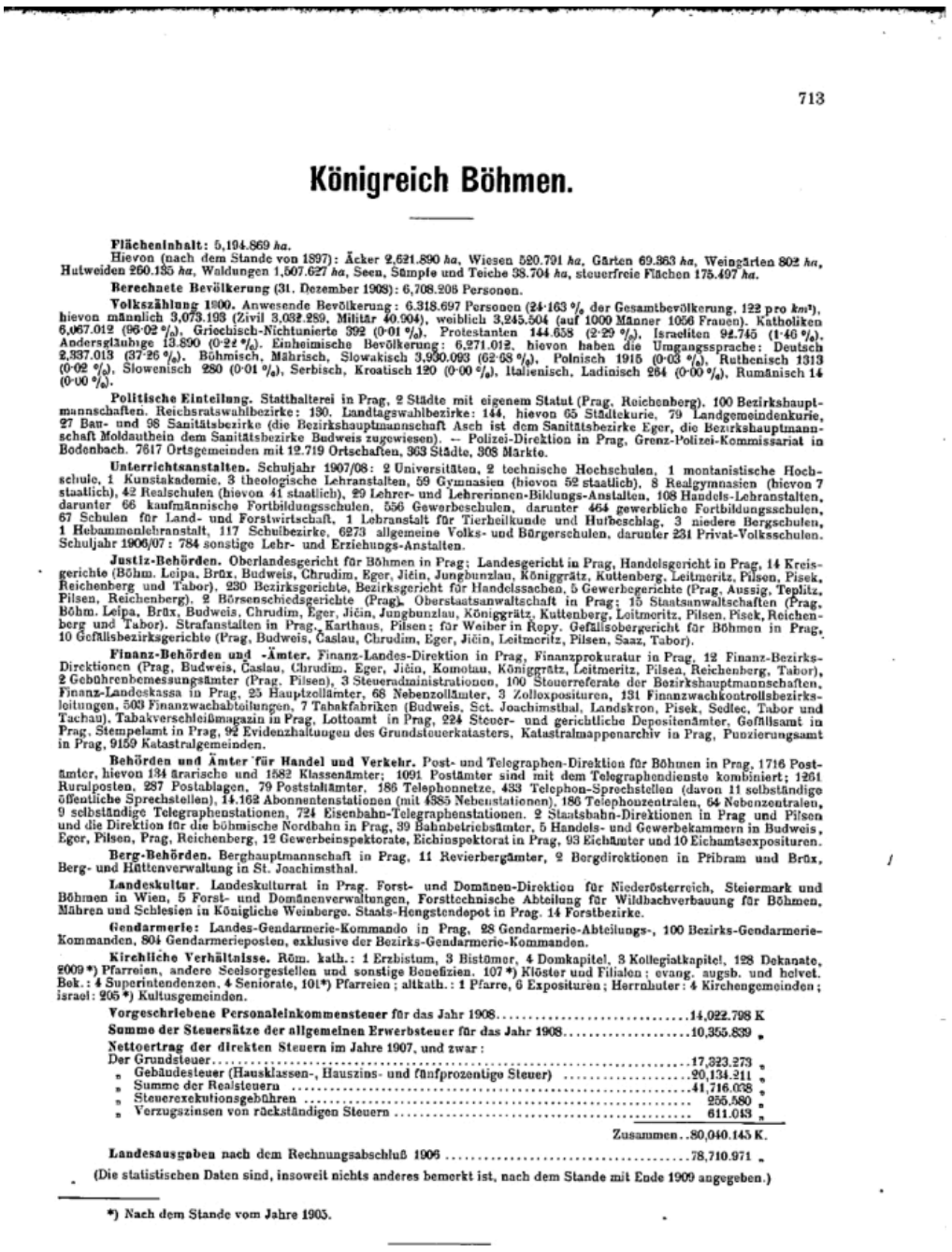


Рис.1.3 - Схема обробки елементів макету, що ілюструє процес сегментації сторінки та подальшого застосування OCR.

Незважаючи на успіхи, метод має обмеження, зокрема недостатню ефективність для нетипових макетів, як показано на прикладі документа з нетиповим форматуванням (рис.1.4), через обмежену кількість таких прикладів у навчальних даних, а також відсутність деяких символів у кастомізованому шрифті.

Обмеження традиційних методів, таких як висока трудомісткість ручної транскрипції чи низька точність стандартного OCR, підкреслюють необхідність інтеграції сучасних методів та технологій, таких як машинне навчання, у обробку історичних текстів. Запропонований метод є гнучким і може бути адаптований до інших старовинних джерел, що підтверджує його актуальність для гуманітарних досліджень. Таким чином, дослідження Флейшхакера та співавторів обґрунтовує необхідність використання інноваційних підходів для ефективного аналізу історичних документів.



713

Königreich Böhmen.

Flächeninhalt: 5,194.869 *ha*.
Hievon (nach dem Stande von 1897): Äcker 2,621.890 *ha*, Wiesen 520.791 *ha*, Gärten 69.363 *ha*, Weingärten 802 *ha*, Hutweiden 260.135 *ha*, Waldungen 1,507.627 *ha*, Seen, Sümpfe und Teiche 38.704 *ha*, steuerfreie Flächen 175.497 *ha*.

Berechnete Bevölkerung (31. Dezember 1908): 6,708.206 Personen.

Volkszählung 1900. Anwesende Bevölkerung: 6.318.697 Personen (24·163 % der Gesamtbevölkerung, 122 pro *km*²), hievon männlich 3,073.193 (Zivil 3,032.289, Militär 40.904), weiblich 3,245.504 (auf 1000 Männer 1056 Frauen). Katholiken 6,067.012 (96·02 %), Griechisch-Nichtunierte 392 (0·01 %). Protestanten 144.658 (2·29 %), Israeliten 92.745 (1·46 %), Andersgläubige 13.890 (0·22 %). Einheimische Bevölkerung: 6,271.012, hievon haben die Umgangssprache: Deutsch 2,337.013 (37·26 %), Böhmisch, Mährisch, Slowakisch 3,920.093 (62·53 %), Polnisch 1915 (0·03 %), Ruthenisch 1313 (0·02 %), Slowenisch 280 (0·01 %), Serbisch, Kroatisch 120 (0·00 %), Italienisch, Ladinisch 264 (0·00 %), Rumänisch 14 (0·00 %).

Politische Einteilung. Statthalterei in Prag, 2 Städte mit eigenem Statut (Prag, Reichenberg), 100 Bezirkshauptmannschaften. Reichsratswahlbezirke: 130. Landtagswahlbezirke: 144, hievon 65 Städtekurie, 79 Landgemeindenkurie, 27 Bau- und 98 Sanitätsbezirke (die Bezirkshauptmannschaft Asch ist dem Sanitätsbezirke Eger, die Bezirkshauptmannschaft Moldauthein dem Sanitätsbezirke Budweis zugewiesen). — Polizei-Direktion in Prag, Grenz-Polizei-Kommissariat in Bodenbach. 7617 Ortsgemeinden mit 12.719 Ortschaften, 363 Städte, 308 Märkte.

Unterrichtsanstalten. Schuljahr 1907/08: 2 Universitäten, 2 technische Hochschulen, 1 montanistische Hochschule, 1 Kunstakademie, 3 theologische Lehranstalten, 59 Gymnasien (hievon 52 staatlich), 8 Realgymnasien (hievon 7 staatlich), 42 Realschulen (hievon 41 staatlich), 29 Lehrer- und Lehrerinnen-Bildungs-Anstalten, 108 Handels-Lehranstalten, darunter 66 kaufmännische Fortbildungsschulen, 556 Gewerbeschulen, darunter 464 gewerbliche Fortbildungsschulen, 67 Schulen für Land- und Forstwirtschaft, 1 Lehranstalt für Tierheilkunde und Hufbeschlag, 3 niedere Bergschulen, 1 Hebammenlehranstalt, 117 Schulbezirke, 6273 allgemeine Volks- und Bürgerschulen, darunter 231 Privat-Volksschulen. Schuljahr 1906/07: 784 sonstige Lehr- und Erziehungs-Anstalten.

Justiz-Behörden. Oberlandesgericht für Böhmen in Prag; Landesgericht in Prag, Handelsgericht in Prag, 14 Kreisgerichte (Böhm. Leipa, Brüx, Budweis, Chrudim, Eger, Jičin, Jungbunzlau, Königgrätz, Kuttenberg, Leitmeritz, Pilsen, Pisek, Reichenberg und Tabor), 230 Bezirksgerichte, Bezirksgericht für Handelssachen, 5 Gewerbegerichte (Prag, Aussig, Teplitz, Pilsen, Reichenberg). 2 Börsenschiedsgerichte (Prag). Oberstaatsanwaltschaft in Prag; 15 Staatsanwaltschaften (Prag, Böhm. Leipa, Brüx, Budweis, Chrudim, Eger, Jičin, Jungbunzlau, Königgrätz, Kuttenberg, Leitmeritz, Pilsen, Pisek, Reichenberg und Tabor). Strafanstalten in Prag, Karthaus, Pilsen; für Weiber in Repy. Gefällsobergericht für Böhmen in Prag, 10 Gefällsbezirksgerichte (Prag, Budweis, Caslau, Chrudim, Eger, Jičin, Leitmeritz, Pilsen, Saaz, Tabor).

Finanz-Behörden und -Ämter. Finanz-Landes-Direktion in Prag, Finanzprokuratur in Prag, 12 Finanz-Bezirks-Direktionen (Prag, Budweis, Caslau, Chrudim, Eger, Jičin, Komotau, Königgrätz, Leitmeritz, Pilsen, Reichenberg, Tabor), 2 Gebührenbemessungsämter (Prag, Pilsen), 3 Steueradministrationen, 100 Steuerreferate der Bezirkshauptmannschaften, Finanz-Landeskassa in Prag, 25 Hauptzollämter, 68 Nebenzollämter, 3 Zollexposituren, 131 Finanzwachkontrollsbezirksleitungen, 503 Finanzwachabteilungen, 7 Tabakfabriken (Budweis, Sct. Joachimsthal, Landskron, Pisek, Sedlec, Tabor und Tachau), Tabakverschleißmagazin in Prag, Lottoamt in Prag, 224 Steuer- und gerichtliche Depositenämter, Gefällsamt in Prag, Stempelamt in Prag, 92 Evidenzhaltungen des Grundsteuerkatasters, Katastralmappenarchiv in Prag, Punzierungsamt in Prag, 9159 Katastralgemeinden.

Behörden und Ämter für Handel und Verkehr. Post- und Telegraphen-Direktion für Böhmen in Prag, 1716 Postämter, hievon 134 ärarische und 1582 Klassenämter; 1091 Postämter sind mit dem Telegraphendienste kombiniert; 1261 Ruralposten, 287 Postablagen, 79 Poststallämter, 186 Telephonnetze, 433 Telephon-Sprechstellen (davon 11 selbständige öffentliche Sprechstellen), 14.162 Abonnentenstationen (mit 4385 Nebenstationen), 186 Telephonzentralen, 64 Nebenzentralen, 9 selbständige Telegraphenstationen, 724 Eisenbahn-Telegraphenstationen. 2 Staatsbahn-Direktionen in Prag und Pilsen und die Direktion für die böhmische Nordbahn in Prag, 39 Bahnbetriebsämter, 5 Handels- und Gewerbekammern in Budweis, Eger, Pilsen, Prag, Reichenberg, 12 Gewerbeinspektorate, Eichinspektorat in Prag, 93 Eichämter und 10 Eichamtsexposituren.

Berg-Behörden. Berghauptmannschaft in Prag, 11 Revierbergämter, 2 Bergdirektionen in Přibram und Brüx, Berg- und Hüttenverwaltung in St. Joachimsthal.

Landeskultur. Landeskulturrat in Prag. Forst- und Domänen-Direktion für Niederösterreich, Steiermark und Böhmen in Wien, 5 Forst- und Domänenverwaltungen, Forsttechnische Abteilung für Wildbachverbauung für Böhmen, Mähren und Schlesien in Königliche Weinberge. Staats-Hengstendepot in Prag. 14 Forstbezirke.

Gendarmerie: Landes-Gendarmerie-Kommando in Prag, 28 Gendarmerie-Abteilungs-, 100 Bezirks-Gendarmerie-Kommanden, 804 Gendarmerieposten, exklusive der Bezirks-Gendarmerie-Kommanden.

Kirchliche Verhältnisse. Röm. kath.: 1 Erzbistum, 3 Bistümer, 4 Domkapitel, 3 Kollegiatkapitel, 128 Dekanate, 2009 *) Pfarreien, andere Seelsorgestellen und sonstige Benefizien. 107 *) Klöster und Filialen; evang. augsb. und helvet. Bek.: 4 Superintendenzen, 4 Seniorate, 101*) Pfarreien; altkath.: 1 Pfarre, 6 Expositurен; Herrnhuter: 4 Kirchengemeinden; israel: 205 *) Kultusgemeinden.

Vorgeschriebene Personaleinkommensteuer für das Jahr 1908	14,022.798 K
Summe der Steuersätze der allgemeinen Erwerbsteuer für das Jahr 1908	10,355.839 „
Nettoertrag der direkten Steuern im Jahre 1907, und zwar:	
Der Grundsteuer	17,323.273 „
„ Gebäudesteuer (Hausklassen-, Hauszins- und fünfprozentige Steuer)	20,134.211 „
„ Summe der Realsteuern	41,716.038 „
„ Steuerexekutionsgebühren	255.580 „
„ Verzugszinsen von rückständigen Steuern	611.043 „
Zusammen	80,040.145 K.
Landesausgaben nach dem Rechnungsabschluß 1906	78,710.971 „

(Die statistischen Daten sind, insoweit nichts anderes bemerkt ist, nach dem Stande mit Ende 1909 angegeben.)

*) Nach dem Stande vom Jahre 1905.

Рис.1.4 - Приклад документа “Schematismus” з нетиповим макетом, що спричинив труднощі.

1.2 Поняття літопису як джерела інформації

Літописи є одними з найважливіших історичних джерел, що надають хронологічно впорядковану інформацію про політичні, релігійні, соціальні та культурні події минулого. Вони створювалися монахами, придворними чи іншими освіченими особами, часто відображаючи упередженість авторів або їхніх покровителів. У контексті середньовічних досліджень літописи, є цінними

джерелами, але їхнє вивчення ускладнене через фізичну деградацію, розпорошеність рукописів і потребу в спеціалізованих знаннях. Сучасні цифрові технології, як показано в дослідженнях Едвардса (2018)[2] і Тельчар та співавторів (2024)[15], значно полегшують доступ до літописів і їхній аналіз, що робить їх актуальними для історичних досліджень.

Літописи поєднують факти, інтерпретації та іноді легенди, що робить їх багатогранними, але складними джерелами. Едвардс (2018)[2] зазначає, що середньовічні англійські рукописи, включно з літописами, розпорошені по бібліотеках світу через історичні переміщення, купівлю чи розрізання, як у випадку фрагментів, розрізаних колекціонером Отто Еге. Це ускладнює їхнє вивчення без цифрових архівів, які дозволяють віртуально об’єднувати фрагменти. Крім того, літописи є не лише текстами, а й матеріальними об’єктами, що містять кодикологічні дані (тип паперу, чорнила, палітурки), які доповнюють історичну інформацію. Тельчар та співавтори (2024)[15] підкреслюють фізичні проблеми рукописів, такі як вицвітання чорнила, кислотне пошкодження паперу чи деградація, що ускладнює читання текстів літописів без спеціалізованих методів.

Традиційні методи роботи з літописами, такі як ручна транскрипція, є трудомісткими та потребують глибоких знань палеографії та кодикології. Едвардс (2018)[2] вказує, що без таких знань навіть цифрові копії літописів можуть залишатися лише “дорогим свідченням відсутності думки”, оскільки технології не замінюють розуміння контексту. Цифрові архіви, однак, революціонізують доступ до літописів. Наприклад, оцифровані фонди Британської бібліотеки чи бібліотеки Корпус-Крісті дозволяють дослідникам аналізувати високоякісні зображення без фізичного контакту з крихкими оригіналами, що особливо важливо для збереження літописів.

Тельчар та співавтори (2024)[15] пропонують автоматизовану транскрипцію літописів за допомогою технологій розпізнавання рукописного тексту (HTR), які використовують машинне навчання для створення цифрових текстових версій. Їхній підхід передбачає індивідуальні схеми транскрипції, адаптовані до дослідницьких питань, наприклад, аналізу орфографічних варіацій чи скорочень у

латинських текстах. Це дозволяє не лише відтворювати текст літопису, а й досліджувати внесок писарів, що додає цінності історичному аналізу. Наприклад, збереження скорочень може вказувати на регіональні чи хронологічні особливості створення літопису.

Незважаючи на переваги, цифрові методи мають обмеження. Едвардс (2018)[2] попереджає, що цифрові зображення можуть не передавати матеріальних характеристик літописів, таких як текстура паперу чи сліди використання, що є ключовими для інтерпретації. Тельчар та співавтори (2024)[15] зазначають, що універсальні HTR-моделі можуть бути неефективними для унікальних шрифтів чи скорочень літописів, що вимагає спеціалізованого налаштування. Крім того, вибір рукописів для оцифрування часто залежить від фінансових чи прагматичних факторів, що може обмежувати доступ до менш відомих літописів.

Літописи залишаються незамінними джерелами для історичних досліджень, а цифрові технології значно розширюють їхній потенціал. Цифрові архіви, як показано Едвардсом (2018)[2], забезпечують глобальний доступ, а HTR-методи, описані Тельчар та співавторами (2024)[15], дозволяють автоматизувати транскрипцію та проводити масштабний аналіз. Однак ефективність цих технологій залежить від поєднання з традиційними гуманітарними знаннями. Таким чином, літописи як джерела інформації підтверджують необхідність інтеграції цифрових підходів у дослідження історичних текстів, що є актуальним для цієї дипломної роботи.

1.3 Основи оптичного розпізнавання символів (OCR)

Оптичне розпізнавання символів (Optical Character Recognition, OCR) — це технологія, що забезпечує автоматичне перетворення зображень тексту, отриманих шляхом сканування чи фотографування, у машинно-зчитуваний формат. OCR відіграє ключову роль у цифровій обробці документів, зокрема історичних, дозволяючи автоматизувати введення даних і полегшувати доступ до текстової інформації. У своїй главі “Historical Review of Theory and Practice of Handwritten Character Recognition” Морі[14] описує основи OCR, фокусуючись на проблемах

розпізнавання символів, рівнях їхньої складності та підходах до вирішення, що є актуальним для розуміння технології в контексті історичних досліджень.

Проблеми OCR можна класифікувати за рівнями складності, які відображають еволюцію технології та її виклики. Виділяють чотири основні рівні (0–3), що базуються на варіативності форм символів, ступені шуму та проблемі сегментації. Рівень 0 стосується розпізнавання друкованих символів одного шрифту з фіксованим розміром і мінімальним шумом, що було основною задачею до середини XX століття і наразі повністю вирішено комерційними OCR-системами. Рівень 1 охоплює символи з невеликими геометричними варіаціями та шумом, тоді як рівень 2 включає більш складні випадки, такі як розпізнавання різних шрифтів чи деградованих текстів. Найвищий рівень 3 пов’язаний із проблемою сегментації, коли символи, особливо в рукописному тексті, не можуть бути легко відокремлені, як у випадку курсивного письма. Ця проблема є центральною для обробки історичних рукописів, де текст часто неструктурований або містить злиті символи.

OCR-процес передбачає кілька етапів: попередню обробку зображення для видалення шуму та корекції нахилу, сегментацію для виділення окремих символів або слів і власне розпізнавання шляхом порівняння з еталонними шаблонами чи моделями. Морі[14] акцентує на важливості сегментації, особливо для рукописних текстів, де відсутність чітких меж між символами значно ускладнює обробку. Наприклад, у японській поштовій системі для полегшення сегментації використовувалися спеціальні поля для написання цифр, що ілюструє необхідність компромісу між людиною та машиною в OCR.

В джерелі[14] виділяють два основні підходи до розпізнавання символів: функціональний (просторовий) аналіз і структурний (описовий) аналіз. Функціональний аналіз базується на математичній структурі векторних просторів із внутрішніми добутками, що ближче до статистичних методів, і ефективний для стандартних шрифтів. Структурний аналіз, навпаки, фокусується на описі геометричних чи топологічних характеристик символів, що краще підходить для рукописних текстів із високою варіативністю. Обидва підходи мають свої сильні

та слабкі сторони, але їхнє поєднання є перспективним для обробки складних документів, таких як історичні рукописи.

Виклики OCR включають високу варіативність форм символів, шум у зображеннях і складність сегментації, особливо для курсивного письма чи деградованих текстів. Зазначається, що навіть людське розпізнавання рукописного тексту досягає 100% лише за умови акуратного письма, тоді як сучасні OCR-системи мають значно нижчу точність у таких задачах.[14] Для історичних документів ці проблеми посилюються через фізичне зношення, нестандартні шрифти та нерівномірне освітлення, що вимагає адаптивних алгоритмів.

OCR є незамінним інструментом для цифрових гуманітарних наук, дозволяючи оцифровувати історичні тексти та полегшувати їхній аналіз. Успіх OCR залежить від чіткого визначення задачі, якості сегментації та вибору відповідного підходу до розпізнавання. Ці основи залишаються актуальними для сучасних систем, які використовують машинне навчання, але базуються на принципах, сформульованих у ранніх дослідженнях. Таким чином, розуміння основ OCR є необхідним для розробки ефективних методів обробки історичних текстів у рамках цієї дипломної роботи.

1.4 Особливості розпізнавання історичних шрифтів

Розпізнавання історичних шрифтів, таких як Fraktur, Antiqua та інші латинські шрифти, є складним завданням у межах оптичного розпізнавання символів (OCR) через їхню високу варіативність, нестандартні форми, складні макети та деградацію документів. Історичні тексти, що охоплюють період від XV до XIX століття, містять різноманітні типографські стилі, лігатури, декоративні елементи та застарілу орфографію, що ускладнює автоматизовану обробку. Спираючись на дослідження Мартінека, Ленца та Краля (2020)[10] та сучасні підходи до поліфонтових моделей OCR, розглянемо ключові особливості розпізнавання історичних шрифтів і методи їхньої обробки.

Історичні шрифти, такі як Fraktur, використаний у німецьких газетах “Ascher Zeitung” XIX століття, характеризуються складними графічними формами та неоднорідністю, оскільки документи можуть містити суміш Fraktur і латинських

шрифтів із різними стилями (рис.1.5). Зазнається, що такі шрифти ускладнюють сегментацію через нерівномірні макети, наприклад двоколонкові структури, та шум, спричинений вицвітанням чорнила чи фізичними пошкодженнями. Для вирішення цих проблем автори застосовують повністю згорткові нейронні мережі (FCN), зокрема U-Net, для сегментації текстових блоків і методи ARU-Net та Kraken для сегментації рядків, що дозволяють адаптуватися до складних макетів і декоративних елементів Fraktur.

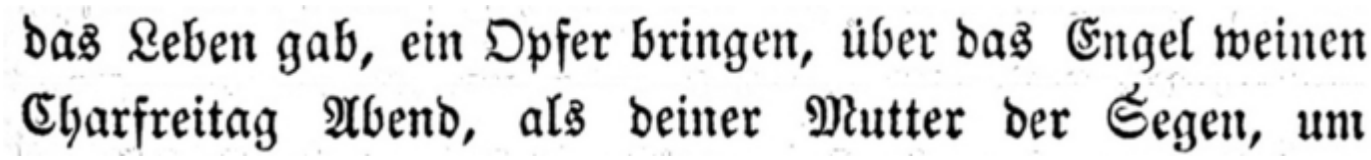


Рис.1.5 - Приклад шрифту Fraktur, що ілюструє складні графічні форми та варіативність історичних шрифтів.

Сучасні підходи до розпізнавання історичних шрифтів включають створення поліфонтових моделей, які здатні обробляти різноманітні шрифти (Fraktur, Antiqua) та мови (німецька, латинська, французька) з Character Error Rate (CER) близько 2% без додаткового налаштування. Такі моделі тренуються на широкому наборі даних, що охоплюють тексти з XV до XIX століття, використовуючи комбінацію попереднього навчання, аугментації даних і голосування. Попередня обробка, як-от бінаризація та корекція повороту, збагачує тренувальні дані, підвищуючи стійкість моделей. Двоетапний підхід, що передбачає навчання на всіх доступних, часто незбалансованих даних, з подальшим тонким налаштуванням на збалансованому піднаборі, дозволяє досягти CER 1.73% на 29 книгах, що на 40% краще за стандартні моделі з CER 2.84%. Тонке налаштування для специфічних класів документів, наприклад ранньомодерних латинських текстів, знижує CER до 1.47%, що на 50% ефективніше за навчання з нуля та на 30% краще за стандартні моделі.

Дане дослідження[10] пропонує метод, що поєднує згорткові (CNN) і рекурентні нейронні мережі (RNN) із Connectionist Temporal Classification (CTC), дозволяючи обробляти цілі рядки Fraktur без сегментації на символи. Цей підхід, проілюстрований на прикладі CTC-вирівнювання (рис.1.6), враховує контекст,

значно зменшуючи помилки порівняно з традиційними методами, які сегментують окремі символи, і забезпечує вищу точність для складних шрифтів. Обмежена кількість анотованих даних компенсується синтетичними даними типу “Hybrid” (зображення символів із випадковими пробілами), які краще імітують Fraktur, ніж “Generated” дані, досягаючи CER 2.2% після тонкого налаштування на 10 сторінках реальних даних. Порівняння з Tesseract (CER 4.5–5.3%) і Transkribus (CER 2.7%) показує, що запропонований метод є ефективнішим, досягаючи CER 2.4% на даних Porta fontium.

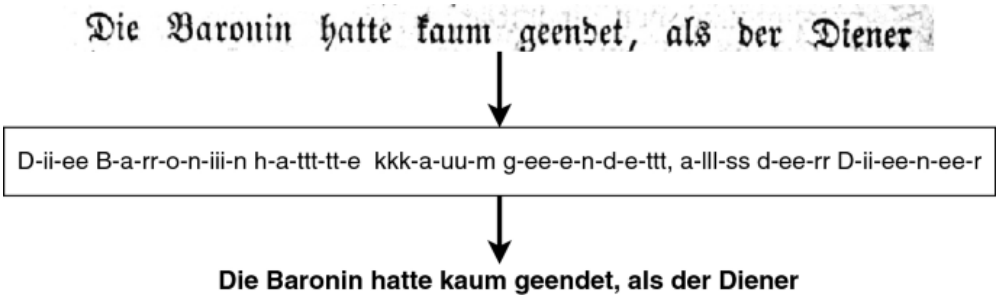


Рис.1.6 - Приклад CTC-вирівнювання для тексту Fraktur, що демонструє перевагу контекстного розпізнавання рядків над традиційною сегментацією.

Виклики розпізнавання історичних шрифтів включають їхню варіативність, застарілу орфографію та деградацію документів. Поліфонтові моделі та синтетичні дані дозволяють створювати універсальні OCR-системи, які можна тонко налаштовувати для конкретних шрифтів чи епох із мінімальними зусиллями. Дослідження Мартінека та співавторів (2020)[10] і сучасні підходи до поліфонтових моделей підтверджують, що комбінація FCN, RNN із CTC і синтетичних даних є ефективною для обробки історичних шрифтів, таких як Fraktur.

1.5 Обробка природної мови (NLP): поняття і застосування

Обробка природної мови (Natural Language Processing, NLP) — це міждисциплінарна галузь, що поєднує комп’ютерні науки, штучний інтелект і лінгвістику для автоматичного аналізу, розуміння та генерації людської мови. NLP охоплює такі завдання, як токенізація, морфологічне тегування, синтаксичний парсинг, класифікація документів, аналіз сентименту та самаризація, які

дозволяють комп’ютерам обробляти текстові дані для інформаційного пошуку, вилучення знань і автоматизації аналізу. У контексті історичних досліджень і аналізу довгих документів NLP відіграє ключову роль, забезпечуючи доступ до складних текстів із нестандартною орфографією чи великим обсягом. Спираючись на дослідження Петтерссон, Мегесі та Нівре (2013)[7] про нормалізацію історичних текстів і сучасні підходи до аналізу довгих документів за допомогою глибоких нейронних мереж (DNN), розглянемо поняття NLP та її застосування.

NLP передбачає розробку алгоритмів для обробки структурованих і неструктурованих текстів, включаючи історичні документи та довгі тексти. Дослідження підкреслюють, що історичні шведські тексти XVI-XIX століть, такі як судові записи та церковні документи, ускладнені відсутністю орфографічних стандартів, фонетичним написанням і впливом інших мов, наприклад німецької. Для адаптації сучасних NLP-інструментів, тренованих на сучасних корпусах, застосовують нормалізацію - перетворення історичного правопису на сучасний. Вони розробили 29 ручних правил нормалізації на основі судових записів 1638 року та реформи правопису 1906 року, які замінюють застарілі літери (наприклад, q → k у qvarn → kvarn) або видаляють надлишкові (hvar → var). Ці правила, проілюстровані в таблиці найпоширеніших змін правопису (табл. 1.1), виявилися ефективними для текстів різних періодів і жанрів, підвищуючи точність тегування дієслів із 66.3% до 78.8% за показником recall і зменшуючи помилки парсингу додатків із 27.4% до 22.7%.[7].

Таблиця 1.1

Найпоширеніших змін правопису для нормалізації шведських історичних текстів XVI–XVIII століть.

16 cent	17 cent	18 cent	Court	Church
-h	-h	w/v	-h	-h
w/v	-dbl	-h	-dbl	w/v
e/ä	w/v	+dbl	w/v	-dbl
-f	-f	-e	-e	e/ä

-dbl	-e	-f	-f	-f
e/a	e/a	e/ä	+dbl	e/ä
-e	+dbl	f/v	e/ä	-e
+dbl	e/ä	e/a	e/a	+dbl
u/v	f/v	-dbl	-i	c/k
c/k	i/j	c/k	f/v	i/j

Сучасні досягнення в NLP, зокрема впровадження глибоких нейронних мереж (DNN), значно розширили можливості аналізу довгих документів, таких як юридичні тексти, медичні записи, фінансові звіти чи історичні архіви. DNN, включаючи рекурентні нейронні мережі (RNN) і трансформери, дозволяють ефективно обробляти великі обсяги тексту, вирішуючи завдання класифікації документів, самаризації та аналізу сентименту. Наприклад, у задачах веб-майнінгу, агрегації новин чи оцінки екологічного впливу DNN забезпечують автоматизоване розуміння текстів, що раніше вимагало значних людських зусиль. Однак аналіз довгих документів відрізняється від обробки коротких текстів через потребу в моделюванні довгих залежностей і управлінні великими контекстами, що вимагає спеціалізованих алгоритмів і значних обчислювальних ресурсів.

Застосування NLP до історичних текстів і довгих документів має спільні виклики, зокрема варіативність мови та обмежену кількість анотованих даних. Дослідження Петтерссона та співавторів (2013)[7] вирішує проблему історичних текстів шляхом нормалізації, що дозволяє використовувати тегер HunPOS і парсер, треновані на сучасному корпусі Stockholm-Umeå Corpus, для вилучення вербальних конструкцій із судових записів. Це підвищує ефективність аналізу економічних і соціальних аспектів ранньомодерного періоду. Водночас сучасні DNN-моделі для довгих документів, як описано в огляді, використовують передові методи, такі як попереднє навчання на великих корпусах і тонке налаштування, для адаптації до специфічних доменів, що може бути застосовано до історичних текстів для масштабного аналізу. Обмеженнями залишаються трудомісткість створення правил нормалізації, граматичні відмінності, які не вирішує

нормалізація, та потреба в оптимізації DNN для обробки великих текстів без втрати точності.

NLP є незамінним інструментом для автоматизації аналізу історичних текстів і довгих документів, полегшуючи доступ до інформації в гуманітарних і прикладних дослідженнях. Дослідження[7] демонструє ефективність нормалізації для адаптації сучасних NLP-інструментів до історичних текстів, тоді як DNN відкривають нові можливості для обробки великих обсягів даних. Ці підходи обґрунтовують актуальність використання NLP у цій дипломній роботі для аналізу історичних текстів.

1.6 Застосування NLP для обробки архаїчних текстів

Обробка природної мови (Natural Language Processing, NLP) відіграє важливу роль у цифрових гуманітарних дослідженнях, дозволяючи автоматизувати аналіз архаїчних текстів, таких як історичні рукописи, листи чи стародруки. Ці тексти ускладнені нестандартним правописом, варіативністю орфографії, застарілою граматикою та фізичною деградацією документів, що робить їх обробку викликом. NLP пропонує методи нормалізації, тегування частин мови (POS tagging) і контекстного аналізу для полегшення пошуку інформації та вилучення знань. Спираючись на дослідження Робертсона та Голдвотера[1] і Макарова та Клематіда[12], розглянемо ключові застосування NLP для обробки архаїчних текстів.

Нормалізація є центральним NLP-процесом для архаїчних текстів, який перетворює історичні словоформи на сучасні еквіваленти, підвищуючи їхню пошукову доступність і придатність для подальшого аналізу. Робертсон і Голдвотер[1] досліджували два нейронні енкодер-декодерні моделі — із м’якою увагою та жорсткою монотонною увагою — для нормалізації текстів п’ятьма мовами (англійська, німецька, угорська, ісландська, шведська) XIV–XVII століть. Їхні результати, проілюстровані графіком точності нормалізації для небачених слів (рис. 1.8), показують, що нейронні моделі значно перевершують наївний базовий метод (який запам’ятовує найчастіші нормалізації) на небачених словах, досягаючи точності до 65.7% для жорсткої уваги порівняно з 2.9–41.4% для

базового методу. Однак при застосуванні до POS-тегування англійських листів XV–XVII століть нормалізація не завжди покращувала результати порівняно з базовим методом, підкреслюючи важливість оцінки впливу на подальші завдання.

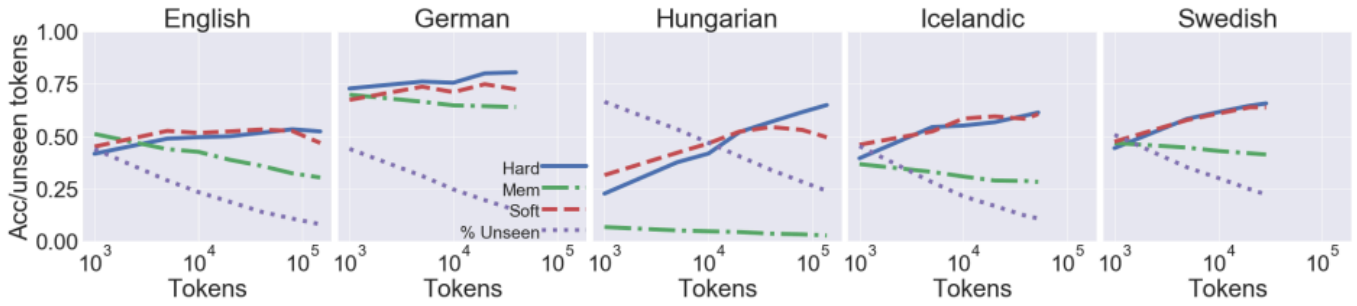


Рис.1.8 - Графік що демонструє перевагу моделей із м’якою та жорсткою увагою над базовим методом.

Напівконтрольована контекстна нормалізація є інноваційним застосуванням NLP, яке використовує немарковані історичні дані для зменшення потреби в анотованих корпусах. Макаров і Клематід[12] розробили генеративну модель на основі шумового каналу, що поєднує нейронний трансд’юсер із мовною моделлю для нормалізації текстів вісьмома мовами, включаючи німецьку, англійську та словенську. Їхній підхід, проілюстрований порівнянням точності для однозначних і неоднозначних слів (рис. 1.9), досягає середньої точності 84.73% при 5000 токенах маркованих даних, що на 8% краще за статистичну машинну перекладну модель (cSMT). У повністю некотрольованому режимі модель забезпечує 88.4% від найкращих контрольованих результатів, демонструючи ефективність використання контексту для розв’язання нормалізаційної неоднозначності, наприклад, у німецькому слові “defz” (нормалізується як “das”, “des” чи “dessen” залежно від контексту).

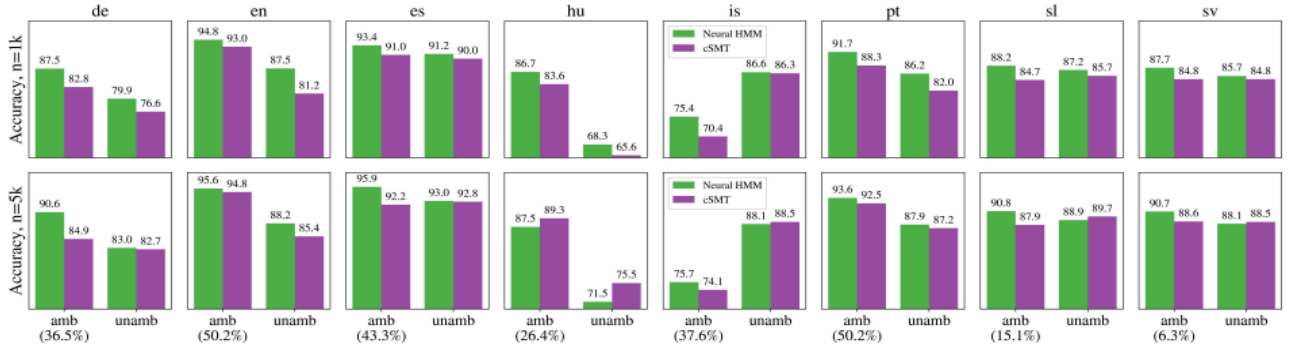


Рис.1.9 - Порівняння точності нормалізації однозначних і неоднозначних, що підкреслює перевагу контекстного підходу для архаїчних текстів.

Застосування NLP до архаїчних текстів стикається з викликами, такими як висока варіативність правопису, низька якість токенізації та обмежена кількість маркованих даних. Робертсон і Голдвотер[1] зазначають, що нейронні моделі можуть недооцінювати бачені слова порівняно з наївним базовим методом, що вимагає гібридних підходів. Макаров і Клематід[12] вказують на залежність від якості токенізації та покриття мовної моделі, особливо для мов із багатою морфологією, як угорська. Обидва дослідження підкреслюють потребу в оцінці як внутрішньої (точність нормалізації), так і зовнішньої (вплив на POS-тегування) ефективності.

NLP значно полегшує обробку архаїчних текстів, автоматизуючи нормалізацію та контекстний аналіз. Обидва дослідження[1,12] демонструють, що нейронні та напівконтрольовані моделі підвищують точність аналізу історичних документів, що є актуальним для цієї дипломної роботи.

2. МЕТОДИ І ЗАСОБИ СТВОРЕННЯ ПРОГРАМНОГО ЗАСТОСУНКУ

2.1 Огляд загальної архітектури системи

Архітектура системи для створення програмного застосунку, призначеного для обробки архаїчних текстів, є основою для забезпечення ефективності, масштабованості та точності аналізу історичних документів. Система побудована на модульному підході, який інтегрує компоненти для оцифрування, попередньої обробки, аналізу даних і зберігання, поєднуючи апаратні засоби з програмними модулями, такими як оптичне розпізнавання символів (OCR) і обробка даних. Спираючись на статтю Revolution Data Systems[13] і дослідження Чакісі та Качді[8], розглянемо ключові компоненти архітектури системи для обробки архаїчних текстів.

Загальна архітектура системи включає три основні рівні: збір і оцифрування даних, обробка та аналіз, зберігання та представлення результатів. На рівні збору даних використовуються апаратні засоби, такі як планетарні сканери, цифрові камери або безпілотні літальні апарати (БПЛА), для створення високоякісних зображень історичних документів, як описано в Revolution Data Systems[13] і Чакісі та Качді[8]. Revolution Data Systems[13] підкреслює важливість вибору обладнання, наприклад, сканерів із підтримкою формату TIFF для архівної якості, що мінімізує пошкодження делікатних рукописів. Чакісі та Качді[8] додають, що фотограмметрія та 3D-лазерне сканування забезпечують точне захоплення даних, як показано в схемі методів документації (рис. 2.1). Попередня обробка включає бінаризацію, корекцію викривлень і сегментацію зображень для підготовки до OCR.

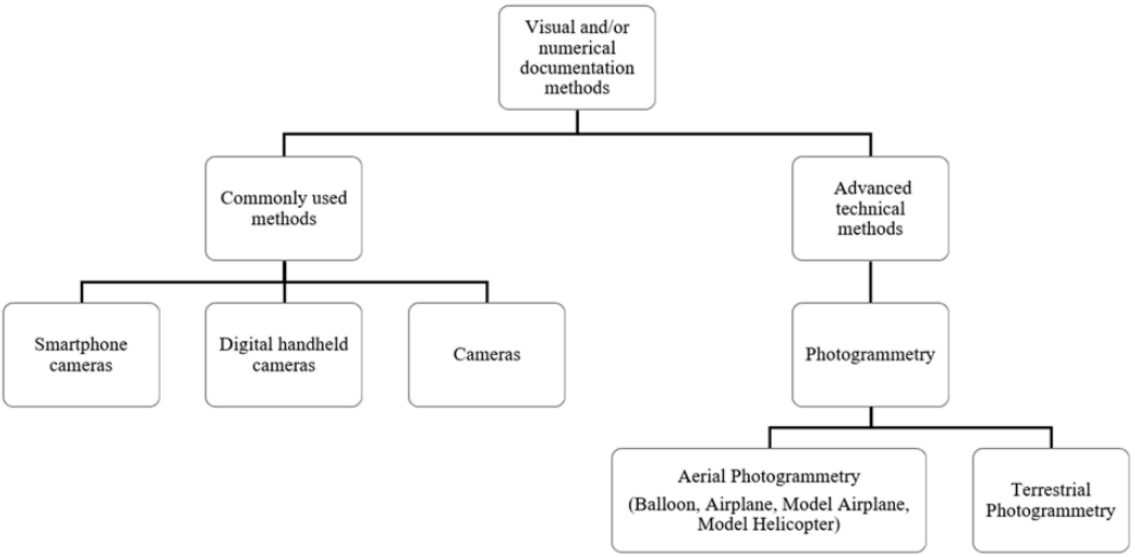


Рис.2.1 - Схема візуальних і числових методів документації, що ілюструє апаратні засоби для збору даних.

Рівень обробки та аналізу охоплює модулі OCR і обробки даних. Модуль OCR, наприклад, Tesseract або ABBYY FineReader, перетворює зображення тексту в машинно-читабельний формат, що є необхідним для архаїчних текстів із нестандартним правописом, таких як Fraktur. Revolution Data Systems[13] зазначає, що OCR потребує постобробки для виправлення помилок, особливо для складних шрифтів, і рекомендує інтеграцію програмного забезпечення для підвищення точності. Чакісі та Качді[8] описують спеціалізоване програмне забезпечення, таке як Photomodeler, Autodesk Recap, Faro Scene і JRC 3D Reconstructor, яке обробляє фотограмметричні дані для створення 2D-чертежів і 3D-моделей, як показано в схемі програм для обробки даних (рис. 2.2). Ці програми можуть бути адаптовані для текстових даних, наприклад, для створення структурованих текстових представлень або пошукових індексів, що є частиною аналітичного модуля.

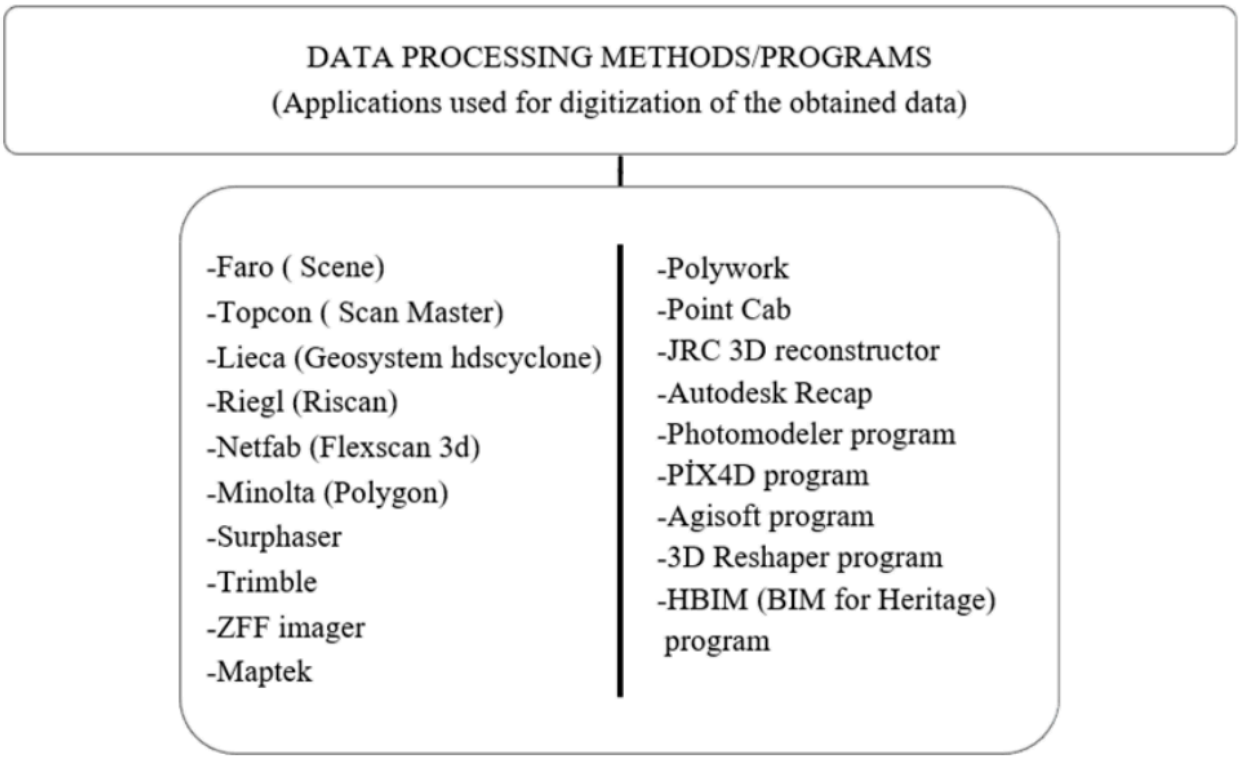


Рис.2.2 - Схема програмного забезпечення для обробки даних.

Рівень зберігання та представлення результатів включає бази даних для збереження оцифрованих текстів, оброблених даних і метаданих, а також інтерфейси для доступу користувачів, наприклад, веб-додатки. Revolution Data Systems[13] наголошує на використанні форматів зберігання, таких як PDF/A, для довготривалого архівування, і систем управління контентом для забезпечення пошуку та доступності. Чакıcı та Качді[8] додають, що результати обробки, такі як 3D-моделі чи ортофотографії, можуть бути представлені через віртуальні тури, що полегшує міждисциплінарний аналіз. Бази даних, наприклад, SQL або NoSQL, підтримують структуроване зберігання, тоді як API дозволяють інтегрувати систему з іншими платформами для обміну даними.

Виклики архітектури включають забезпечення якості оцифрування, обробку низькоякісних сканів і сумісність між апаратними та програмними компонентами. Revolution Data Systems[13] підкреслює необхідність ретельного планування, включаючи вибір обладнання, щоб уникнути помилок оцифрування через крихкість документів. Чакıcı та Качді[8] зазначають, що вибір відповідного програмного забезпечення, наприклад, Faro Scene для лазерного сканування, є

критично важливим для точності обробки даних. Модульна архітектура дозволяє масштабувати систему, додаючи нові компоненти, наприклад, для аналізу текстів чи створення інтерактивних візуалізацій.

Загальна архітектура системи для обробки архаїчних текстів забезпечує інтеграцію оцифрування, обробки та зберігання даних, поєднуючи апаратні засоби з програмними модулями. Стаття Revolution Data Systems[13] і дослідження Чакісі та Качді[8] демонструють важливість модульності та ретельного планування в архітектурі, що є основою для створення програмного застосунку.

2.2 Модуль OCR: вибір і реалізація

Модуль оптичного розпізнавання символів (OCR) є наріжним елементом програмного застосунку для обробки архаїчних текстів, забезпечуючи перетворення зображень історичних документів у машинно-читабельний текстовий формат. Для німецьких текстів у шрифті Fraktur, характерних для документів XV–XIX століть, OCR стикається з викликами через декоративні лігатури, застарілу орфографію та фізичну деградацію паперу. Вибір і реалізація OCR-модуля вимагають урахування точності розпізнавання, адаптивності до історичних шрифтів, мінімальних вимог до анотованих даних і сумісності з модулями обробки природної мови (NLP) для нормалізації та аналізу. Спираючись на дослідження Мартінека, Ленца та Краля[10] і Больльманна[11], розглянемо критерії вибору OCR-системи, методи її реалізації, приклади застосування та виклики.

Вибір OCR-системи залежить від здатності обробляти складні шрифти, як Fraktur, ефективності на низькоякісних сканах, низьких вимог до тренувальних даних і можливості інтеграції з іншими модулями. Мартінека та співавтори[10] розробили OCR-систему на основі глибоких нейронних мереж, що поєднує згорткові (CNN) і рекурентні нейронні мережі (RNN) із Connectionist Temporal Classification (CTC). Ключовою особливістю є використання синтетичних даних, які імітують Fraktur, зменшуючи потребу в маркованих даних. Архітектура системи, проілюстрована схемою нейронної мережі (рис. 2.3), демонструє, як CNN і RNN оптимізують розпізнавання складних шрифтів. Больльманн[11]

зазначає, що якість OCR впливає на нормалізацію, оскільки помилки розпізнавання, як плутанина між “ſ” і “f”, ускладнюють подальшу обробку. Він рекомендує нейронні моделі, такі як Kraken, для Fraktur завдяки їхній здатності адаптуватися до небачених токенів після навчання на різноманітних даних. Оптимальний вибір — нейронні OCR-системи з підтримкою синтетичних даних і контекстної сегментації.

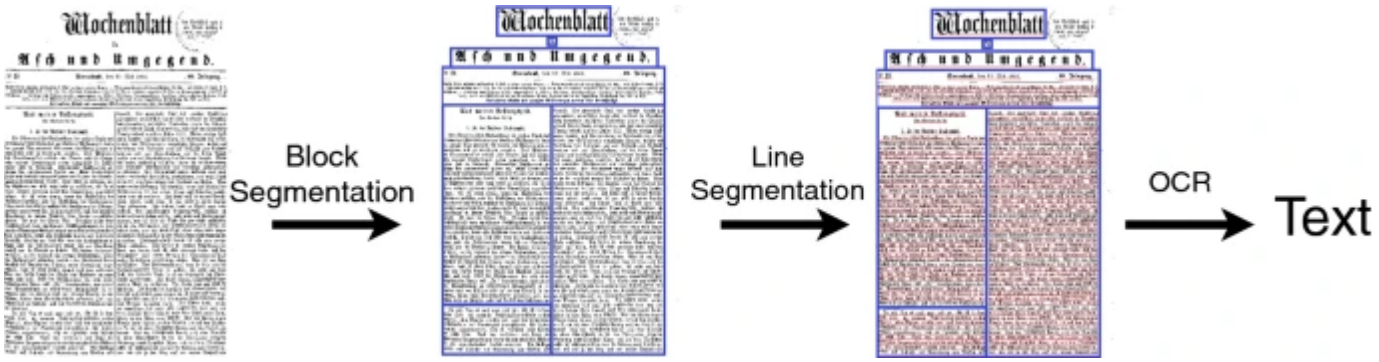


Рис.2.3 - Схема пошагового розпізнавання тексту нейронної OCR-системи.

Реалізація OCR-модуля охоплює кілька етапів: попередню обробку зображень, навчання моделі, сегментацію тексту, інтеграцію з системою та постобробку. Попередня обробка зображень включає бінаризацію для підвищення контрасту, корекцію повороту сторінки та видалення шуму, що необхідно для деградованих документів Fraktur. Навчання моделі передбачає попереднє навчання на синтетичних даних і тонке налаштування на невеликому наборі реальних даних. Мартінека та співавтори[10] використовують U-Net(рис. 2.4) для сегментації текстових блоків і Kraken для виділення рядків, що дозволяє обробляти складні макети, як-от двоколонкові тексти. Сегментація тексту розділяє сторінку на регіони, блоки та рядки, що критично для Fraktur через декоративні елементи. Інтеграція з системою здійснюється через API, яке передає розпізнаний текст до NLP-модуля для нормалізації. Постобробка включає корекцію помилок за допомогою словників або мовних моделей, наприклад, виправлення “ge-schriben” на “geschriben”. Больльманн[11] підкреслює, що кількість тренувальних даних впливає на якість OCR і нормалізації, як показано в кривій навчання для англійського датасету (рис. 2.5), де більший обсяг даних покращує обробку

небачених слів. Реалізація модуля має бути адаптивною, щоб враховувати варіативність Fraktur і низьку якість сканів.

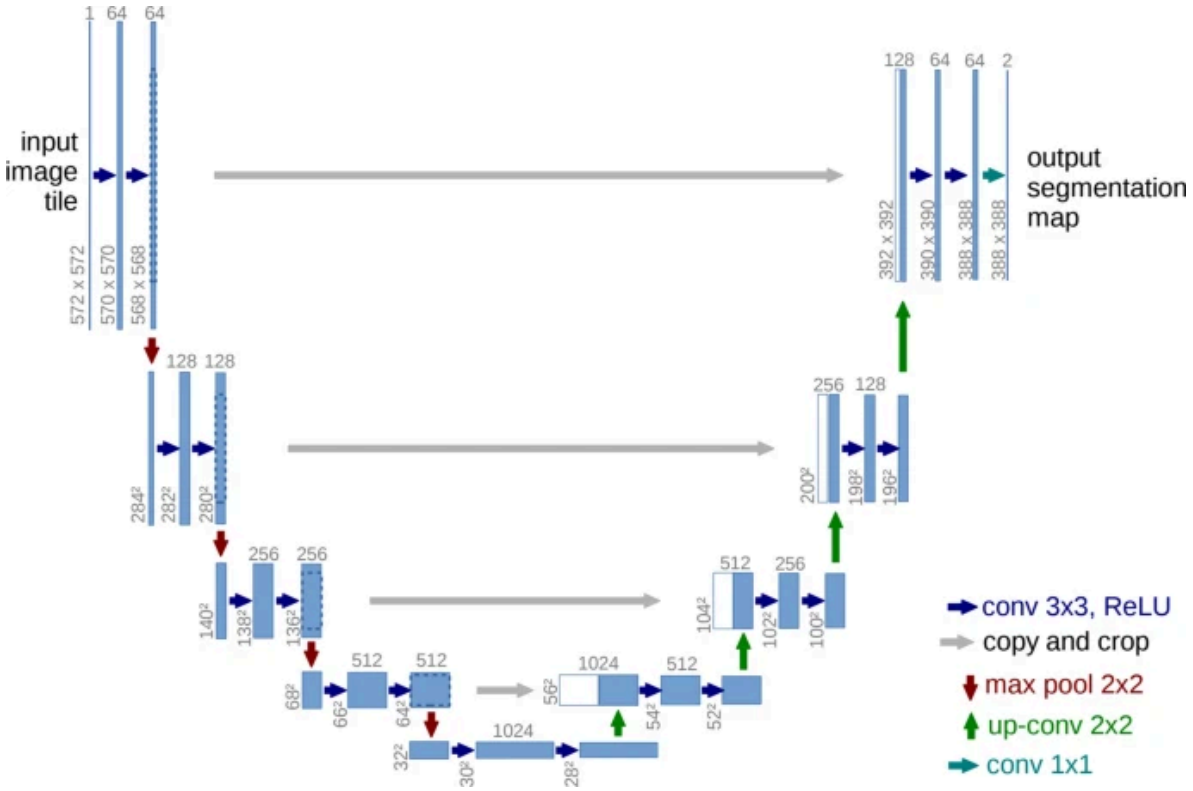


Рис.2.4 - Схема архітектури U-Net.

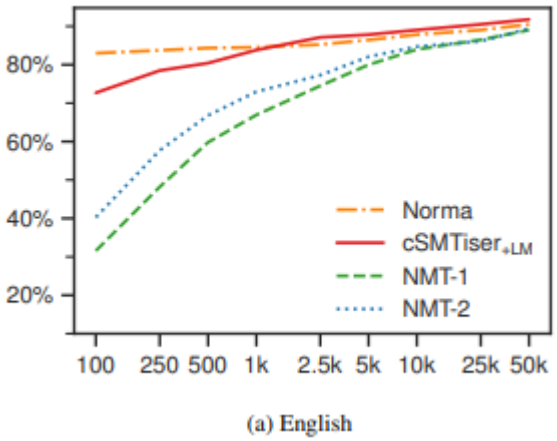


Рис.2.5 - Крива навчання для англійського датасету, що показує залежність точності від кількості тренувальних даних.

Приклади застосування демонструють ефективність OCR для Fraktur. У корпусі Anselm (німецькі релігійні тексти XIV–XVI століть) система Мартінека та співавторів[10] коректно розпізнає “wermuttafft” як “wermutsaft” після виправлення лігатури “ſ”. У корпусі RIDGES (наукові тексти 1482–1652) OCR

обробляє “defz” як “das” у реченні “defz buch”, хоча контекстна неоднозначність вимагає постобробки для вибору між “das”, “des” чи “dessen”. Больльманн[11] зазначає, що для небачених слів, як “tődtet” (вбиває), OCR-системи потребують різноманітних даних, щоб уникнути плутанини з формами, як “todtet”.

Виклики реалізації включають варіативність шрифтів, деградацію документів і обмежену кількість анотованих даних. Мартінека та співавтори[10] вказують, що декоративні елементи Fraktur ускладнюють сегментацію, вимагаючи спеціалізованих моделей, як ARU-Net, для виділення рядків. Больльманн[11] додає, що низька якість сканувань знижує точність OCR до 20% без адаптивної попередньої обробки. Для подолання цих викликів рекомендують синтетичні дані, гібридні підходи (поєднання нейронних моделей із правилами) і постобробку з мовними моделями.

Модуль OCR забезпечує якісне розпізнавання архаїчних текстів Fraktur, створюючи основу для їхньої обробки. Дослідження Мартінека та співавторів[10] і Больльманна[14] демонструють, що нейронні OCR-системи з синтетичними даними та адаптивною обробкою є оптимальними для історичних документів.

2.3 Модуль NLP: попередня обробка, нормалізація, переклад

Модуль обробки природної мови (NLP) є серцем програмного застосунку, призначеного для роботи з архаїчними текстами, зокрема німецькими документами у шрифті Fraktur, які походять із XV–XVIII століть. Ці тексти мають унікальні особливості: нестандартний правопис, декоративні лігатури (наприклад, “ſt” замість “st”), застаріла лексика, а також фізична деградація паперу, що ускладнює їх розпізнавання та аналіз. Модуль NLP вирішує ці виклики через три основні етапи: попередню обробку, яка очищає та структурує текст після оптичного розпізнавання символів (OCR); нормалізацію, що адаптує архаїчні словоформи до сучасних стандартів; і переклад, який забезпечує доступність тексту для сучасних мов або міжмовного аналізу. Ці етапи дозволяють перетворити складні історичні документи на формат, придатний для пошуку, аналізу та досліджень, відкриваючи доступ до культурної спадщини для істориків, лінгвістів і широкої аудиторії. Спираючись на роботи Танга, Капа, Петтерссон та

Нівре[9] і Больманна[11], детально розглянемо кожен етап, їхню реалізацію, приклади та виклики, зосередившись на обробці текстів Fraktur.

Попередня обробка є першим і фундаментальним етапом, який готує сирий текст, отриманий із OCR, до подальшої обробки. Тексти Fraktur, такі як німецькі рукописи чи друковані книги раннього модерного періоду, часто містять помилки розпізнавання через складні шрифти, деградацію паперу або нечіткі сканування. Наприклад, слово “geschriben” (написано) може бути розпізнане як “ge-schriben” через помилкове розділення лігатур або пробілів. Попередня обробка охоплює кілька підетапів: токенізація — розбиття тексту на окремі слова чи фрази; видалення шуму — усунення артефактів OCR, таких як помилкові символи (наприклад, “#” чи “*” через дефекти сканування); виправлення помилок OCR — заміна неправильних символів, як-от “f” на довге “ſ”; і сегментація речень — визначення меж речень у тексті, де пунктуація може бути відсутньою або нестандартною. Танг та співавтори [9] підкреслюють, що для текстів Fraktur токенізація є складною через нестандартні пробіли та лігатури, які збивають із пантелику стандартні алгоритми. Вони пропонують використовувати нейронні моделі машинного перекладу (NMT) із механізмами уваги, які враховують контекст і досягають точності сегментації до 93% для німецьких текстів. Наприклад, у реченні “defz buch ift alt” токенізація має правильно виділити “defz” і “buch” як окремі слова, навіть якщо OCR помилково об’єднує “ift” з “alt”. Їхній аналіз продуктивності, показаний у таблиці результатів нормалізації (табл. 2.1), демонструє, що якісна попередня обробка значно покращує точність наступних етапів. Без належної токенізації та очищення тексту нормалізація може бути неточною, що призведе до помилок у семантичному аналізі чи пошуку.

Таблиця 2.1

Результатів точності нормалізації (Acc) і помилок на рівні символів (CER) для історичних текстів.

	English		**German**		**Hungarian**		**Icelandic**		**Swedish**	
	Acc	CER	Acc	CER	Acc	CER	Acc	CER	Acc	CER
Baseline	94,3	0,07	96,6	0,04	80,1	0,21	84,6	0,19	**92,9**	0,07

NoAtt-RNN	94,73	0,02	94,89	0,02	90,99	0,03	86,73	0,05	91,44	0,03
NoAtt-GRU	94,79	0,02	94,85	0,02	91,03	0,03	86,98	0,05	91,34	0,03
NoAtt-LSTM	94,61	0,02	95,78	0,02	90,91	0,03	86,61	0,05	91,29	0,03
Att-RNN	94,69	0,02	94,23	0,02	91,69	0,02	**87,59**	0,04	91,56	0,03
Att-GRU	94,8	0,02	94,83	0,02	91,68	0,02	87,17	0,05	91,68	0,03
Att-LSTM	94,85	0,02	96	0,02	91,57	0,03	86,83	0,05	91,72	0,03
Transformer	**95,16**	0,02	95,22	0,02	**92,14**	0,02	86,45	0,05	88,99	0,05
BPE-Soft	95,02	0,02	**96,64**	0,01	91,96	0,03	87,19	0,03	91,21	0,03

Нормалізація — це процес приведення архаїчних словоформ до сучасних еквівалентів, що робить текст зрозумілим для сучасних користувачів і придатним для пошуку чи аналізу. У текстах Fraktur, таких як німецькі рукописи чи друковані книги XV–XVII століть, слова часто мають застарілий правопис, який відрізняється від сучасної німецької мови. Наприклад, слово “wermuttafft” (сік полину) у Fraktur потрібно нормалізувати до “wermutsaft”, щоб воно відповідало сучасним стандартам. Больльманн[11] досліджував нормалізацію таких текстів і показав, що нейронні моделі, які враховують контекст, досягають точності до 94%, значно перевершуючи традиційні системи, як Tesseract. Один із яскравих прикладів — слово “defz”, яке залежно від контексту може нормалізуватися як “das” (означений артикль, як у “defz buch” → “das buch”, книга), “des” (родовий артикль) або “dessen” (відносний займенник). Без контексту нормалізація може бути помилковою, що призводить до неправильного тлумачення тексту. Больльманн[11] пропонує використовувати моделі машинного перекладу, які аналізують сусідні слова, щоб визначити правильний варіант. Його порівняння продуктивності систем нормалізації, проілюстроване графіком (рис. 2.6), показує, що якісна попередня обробка (виправлення лігатур і помилок OCR) скорочує помилки нормалізації на 15%. Реалізація нормалізації включає навчання нейронного трансд’юсера на парах історичних і сучасних слів, наприклад, “geschriben” → “geschrieben”, із застосуванням n-грамних мовних моделей для контекстної корекції, щоб врахувати синтаксичні особливості.

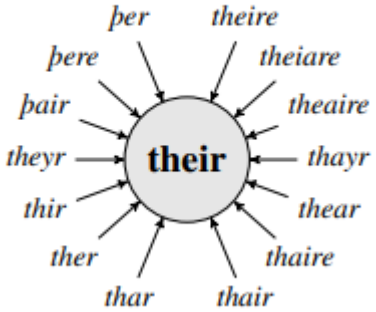


Рис.2.6 - Приклад нормалізації для історичних текстів.

Переклад у контексті архаїчних текстів Fraktur потрібен для адаптації історичного тексту до сучасної мови або для міжмовного аналізу, щоб зробити його доступним для ширшої аудиторії. Наприклад, німецький текст XVII століття у Fraktur може бути перекладений на сучасну німецьку, англійську чи іншу мову для дослідників, які не володіють ранньонововерхньонімецькою. Танг та співавтори [9] пропонують використовувати моделі нейронного машинного перекладу (NMT), такі як Transformer, для перекладу нормалізованих текстів. Наприклад, фраза “wermuttafft in die ohren getropft” спочатку нормалізується до “wermutsaft in die ohren getropft” (сік полину, закапаний у вуха), а потім перекладається на англійську як “wormwood juice dripped in the ears”. Цей процес вимагає не лише нормалізації, а й словників історичних термінів, оскільки слова, як “tődtet” (вбиває), можуть бути відсутніми в сучасних словниках. Больльманн[11] додає, що переклад Fraktur ускладнений синтаксичними особливостями, такими як німецький порядок слів у підрядних реченнях, що потребує попереднього навчання NMT на історичних корпусах. Реалізація перекладу включає інтеграцію з мовними моделями, які враховують контекст, і досягає точності до 86% для міжмовних пар, як-от ранньонововерхньонімецька → сучасна англійська. Наприклад, речення “tődtet die wúrme” нормалізується до “tötet die würmer” і перекладається як “kills the worms”.

Виклики реалізації модуля NLP для текстів Fraktur пов’язані з кількома факторами. По-перше, контекстна неоднозначність: слово “defz” може мати кілька нормалізацій (“das”, “des”, “dessen”), і без аналізу контексту точність нормалізації

падає до 10%, як зазначають Танг та співавтори[9]. По-друге, обмежена кількість анотованих даних: історичні корпуси Fraktur часто мають мало маркованих прикладів, що ускладнює навчання нейронних моделей. По-третє, помилки OCR: лігатури, як “ſt”, і деградація паперу призводять до помилок розпізнавання, які, за Больльманном[11], знижують точність нормалізації на 12% без належної попередньої обробки. Для подолання цих викликів рекомендують: синтетичні дані, які імітують Fraktur для розширення тренувальних наборів; контекстно-чутливі алгоритми, що аналізують сусідні слова; і гібридні підходи, які поєднують нейронні моделі з правило-базованими методами для корекції специфічних помилок, як-от лігатури чи застаріла орфографія.

Модуль NLP забезпечує потужний інструмент для обробки архаїчних текстів Fraktur, дозволяючи очищати, адаптувати та перекладати їх для сучасного використання. Дослідження Танга та співавторів[9] і Больльманна[11] показують, що нейронні моделі з якісною попередньою обробкою та контекстною нормалізацією є оптимальними для історичних німецьких текстів. Ці методи дозволяють не лише зберегти історичну цінність текстів, а й зробити їх доступними для дослідників, істориків і широкої аудиторії через сучасні цифрові платформи.

2.4 Збір та підготовка корпусу літописів

Збір та підготовка корпусу літописів є фундаментальним етапом у розробці програмного застосунку для обробки історичних текстів, зокрема німецьких літописів у шрифті Fraktur. Ці документи, що містять хронологічні записи подій, мають значну історичну, релігійну та правову цінність, але їхній архаїчний правопис, декоративні шрифти та фізична деградація ускладнюють цифрову обробку. Створення якісного корпусу вимагає ретельного відбору репрезентативних текстів, їх оцифрування, попередньої обробки та анотації, щоб забезпечити придатність для аналізу за допомогою технологій OCR та NLP. Німецькі літописи у шрифті Fraktur, характерні для XIV–XIX століть, є цінним джерелом для вивчення історії, культури та мови, але потребують спеціального підходу до збору та підготовки.

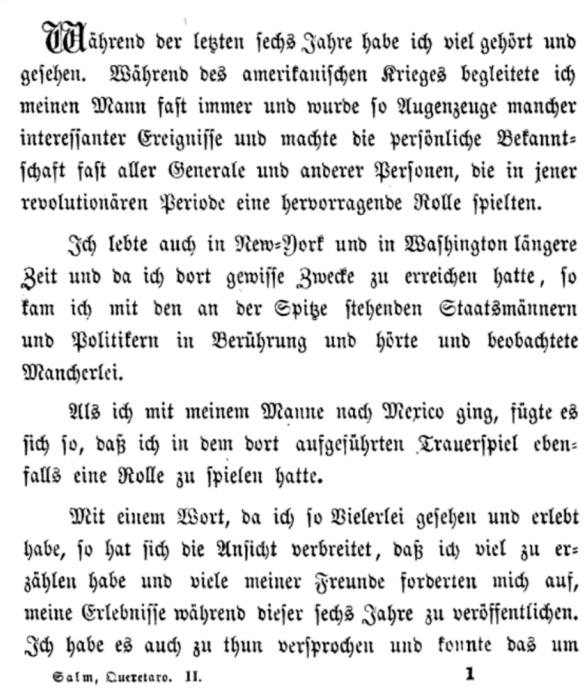
Щоб зібрати корпус німецьких літописів у шрифті Fraktur, необхідно визначити тексти, які відповідають меті дослідження. Літописи, створені в монастирях, містах чи при дворах, документують релігійні, адміністративні чи історичні події. Наприклад, монастирські хроніки, такі як аннали Санкт-Галлена чи Фульди, описують релігійні та політичні події, тоді як міські хроніки Нюрнберга чи Аугсбурга фіксують місцеві події, торгівлю та право. Літописи ранньої нововерхньонімецької доби, датовані 1450–1800 роками, часто друкувалися шрифтом Fraktur і включають записи про події Священної Римської імперії. Важливо відбирати тексти з XIV–XIX століть, написані ранньонововерхньонімецькою або середньоверхньонімецькою, щоб забезпечити мовну однорідність. Корпус має поєднувати релігійні, правові та світські тексти для репрезентативності, а перевага віддається оцифрованим документам із відкритих джерел із ліцензіями Creative Commons або Public Domain, щоб уникнути обмежень використання. Хроніка Нюрнберга 1493 року є прикладом тексту, який відповідає цим критеріям завдяки своїй історичній значущості та доступності в цифрових архівах.

Підготовка корпусу літописів починається зі збору даних із цифрових архівів, бібліотек і репозиторіїв, які надають оцифровані документи. Важливо зберігати метадані, такі як автор, рік створення, місце походження та жанр, щоб полегшити аналіз і категоризацію. Для фізичних документів, які ще не оцифровані, використовуються планетарні сканери для створення високоякісних зображень без пошкодження оригіналів, із збереженням у форматі TIFF для архівної якості. Попередня обробка зображень передбачає бінаризацію для підвищення контрасту, корекцію повороту сторінки та видалення шуму, як-от плям чи потертостей, що є критично важливим для деградованих документів Fraktur. Наступним етапом є застосування нейронних OCR-систем, таких як Kraken або Transkribus, адаптованих до Fraktur, для розпізнавання тексту. Постобробка включає виправлення помилок OCR, наприклад, заміну “ſ” на “s” або корекцію лігатур, як “ſt” → “st”. Анотація тексту додає метадані, такі як нормалізовані форми слів (наприклад, “wermuttafft” → “wermutsaft”), щоб полегшити

NLP-аналіз. Корпус зберігається у структурованому форматі, як TEI XML, із метаданими для сумісності з дослідницькими інструментами, а для управління використовуються бази даних SQL або NoSQL, що забезпечують пошукові можливості.

Відкриті джерела для німецьких літописів у шрифті Fraktur доступні в кількох цифрових репозиторіях. Deutsche Digitale Bibliothek містить оцифровані хроніки, як-от монастирські аннали чи міські хроніки Аугсбурга, із відкритими ліцензіями Creative Commons або Public Domain. Zentrales Verzeichnis Digitalisierter Drucke пропонує доступ до друкованих текстів XVI–XIX століть, включаючи літописи Fraktur, із можливістю завантаження. Bayerische Staatsbibliothek надає оцифровані хроніки Священної Римської імперії у шрифті Fraktur у форматі PDF із відкритим доступом. Europeana Collections містить німецькі літописи Fraktur із метаданими та відкритими ліцензіями, придатними для досліджень. Прикладом літопису є сторінка(рис. 2.7), що ілюструє текст Fraktur.

Підготовка корпусу літописів Fraktur стикається з кількома викликами. Багато літописів доступні лише у фізичних архівах, що вимагає співпраці з бібліотеками чи музеями для оцифрування. Низька якість сканів через деградацію паперу знижує точність OCR до 20% без належної попередньої обробки. Лінгвістична варіативність Fraktur, включаючи архаїчний правопис і лігатури, потребує спеціалізованих моделей OCR і нормалізації. Анотація великих корпусів є трудомісткою, тому синтетичні дані та напів контрольоване навчання стають необхідними. Для подолання цих проблем рекомендується співпрацювати з цифровими архівами, використовувати нейронні OCR-системи, як Kraken, із синтетичними даними, та застосовувати автоматизовані інструменти анотації, як Transkribus. Збір та підготовка корпусу літописів Fraktur створює надійну основу для цифрової обробки, відкриваючи можливості для історичних і лінгвістичних досліджень.



Während der letzten ſechs Jahre habe ich viel gehört und geſehen. Während des amerikaniſchen Krieges begleitete ich meinen Mann faſt immer und wurde ſo Augenzeuge mancher intereſſanter Ereigniſſe und machte die perſönliche Bekanntſchaft faſt aller Generale und anderer Perſonen, die in jener revolutionären Periode eine hervorragende Rolle ſpielten.

Ich lebte auch in New-York und in Waſhington längere Zeit und da ich dort gewiſſe Zwecke zu erreichen hatte, ſo kam ich mit den an der Spitze ſtehenden Staatsmännern und Politikern in Berührung und hörte und beobachtete Mancherlei.

Als ich mit meinem Manne nach Mexico ging, fügte es ſich ſo, daß ich in dem dort aufgeführten Trauerſpiel ebenfalls eine Rolle zu ſpielen hatte.

Mit einem Wort, da ich ſo Vielerlei geſehen und erlebt habe, ſo hat ſich die Anſicht verbreitet, daß ich viel zu erzählen habe und viele meiner Freunde forderten mich auf, meine Erlebniſſe während dieſer ſechs Jahre zu veröffentlichen. Ich habe es auch zu thun verſprochen und konnte das um

Salm, Queretaro. II. 1

Рис.2.7 - Приклад сторінки з текстом Fraktur.

2.5 Інтерфейс користувача та взаємодія з модулями

Інтерфейс користувача (UI) програмного застосунку для обробки архаїчних текстів, зокрема німецьких літописів у шрифті Fraktur, відіграє ключову роль у забезпеченні зручної взаємодії між користувачем і системою. Він слугує посередником, що інтегрує функціональність модулів оптичного розпізнавання символів (OCR) і обробки природної мови (NLP), дозволяючи користувачам завантажувати зображення документів, отримувати розпізнаний текст, нормалізовані форми та переклади. Основна мета інтерфейсу — забезпечити інтуїтивно зрозуміле керування, високу продуктивність і адаптивність до потреб дослідників, істориків і лінгвістів, які працюють із текстами Fraktur. Розглянемо принципи розробки інтерфейсу, його функціональні компоненти та механізми взаємодії з модулями OCR і NLP.

Розробка інтерфейсу користувача базується на принципах простоти, доступності та модульності(рис. 2.8). Для обробки літописів Fraktur інтерфейс має бути веб-орієнтованим, щоб забезпечити кросплатформний доступ через браузери, що усуває потребу в локальній установці. Веб-інтерфейс дозволяє дослідникам із різних регіонів працювати з оцифрованими документами, завантажуючи скановані

зображення. Інтерфейс включає основні компоненти: область завантаження зображень, панель перегляду розпізнаного тексту, секцію для нормалізації та перекладу, а також панель налаштувань для вибору моделей OCR і NLP. Для забезпечення доступності інтерфейс підтримує багатомовність (наприклад, німецьку, англійську) і адаптивний дизайн, що оптимізує відображення на різних пристроях, від настільних комп'ютерів до планшетів. Модульність інтерфейсу дозволяє легко додавати нові функції, як-от аналіз сентименту чи класифікацію текстів, залежно від вимог проекту.

Функціональність інтерфейсу користувача зосереджена на спрощенні взаємодії з модулями OCR і NLP. Користувач завантажує зображення літопису у форматі PNG або JPEG через область завантаження, після чого інтерфейс передає файл до модуля OCR, наприклад, Tesseract, який адаптований до розпізнавання Fraktur. Модуль OCR виконує попередню обробку зображення (бінаризацію, корекцію повороту, видалення шуму) і розпізнає текст, повертаючи сирий результат, як-от "defz buch". Інтерфейс відображає цей текст у панелі перегляду, дозволяючи користувачу перевірити якість розпізнавання та внести ручні правки, наприклад, виправити "ſ" на "s". Далі текст передається до модуля NLP, який нормалізує архаїчні форми (наприклад, "defz" → "das") і, за потреби, перекладає на сучасну мову, як-от англійську ("das buch" → "the book"). Інтерфейс відображає нормалізований і перекладений текст у відповідній секції, забезпечуючи можливість порівняння оригіналу, розпізнаного тексту та результатів обробки. Панель налаштувань дозволяє обрати моделі OCR (наприклад, Tesseract чи Calamari) і NLP (наприклад, Transformer для нормалізації чи deep-translator для перекладу).

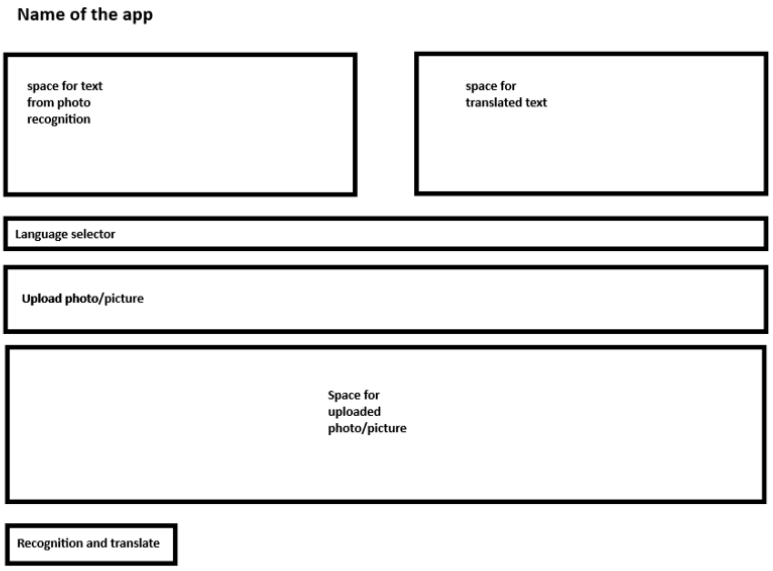


Рис.2.8 - Макет інтерфейсу додатка.

Прикладом взаємодії є обробка сторінки з Нюрнберзької хроніки 1493 року. Користувач завантажує скан через інтерфейс, модуль OCR розпізнає фразу “wermuttafft in die ohren getropft” із виправленням лігатури “ſ”. Модуль NLP нормалізує її до “wermutsaft in die ohren getropft” і перекладає на англійську як “wormwood juice dripped in the ears”. Інтерфейс відображає всі етапи обробки, дозволяючи користувачу перевірити результати та експортувати їх у TEI XML.

Виклики розробки інтерфейсу включають забезпечення високої швидкості обробки великих зображень, адаптацію до різної якості сканів і підтримку користувачів із різним рівнем технічної підготовки. Низька якість сканів може знижувати точність OCR до 20%, що вимагає інтерактивних інструментів для ручної корекції в інтерфейсі. Адаптивний дизайн і багатомовність забезпечують доступність для широкої аудиторії, а модульна структура дозволяє масштабувати застосунок, додаючи нові функції, як-от аналіз метаданих чи візуалізацію текстів. Інтерфейс користувача, інтегрований із модулями OCR і NLP, створює ефективний інструмент для обробки літописів Fraktur, відкриваючи можливості для історичних і лінгвістичних досліджень.

2.6 Застосовані бібліотеки та інструменти

Розробка програмного застосунку для розпізнавання та перекладу німецьких літописів у шрифті Fraktur вимагала використання спеціалізованих бібліотек і інструментів, які забезпечують обробку зображень, оптичне розпізнавання символів (OCR), переклад тексту та створення інтерфейсу користувача. У проекті застосовувалися Tesseract OCR, deep-translator, PIL, Streamlit і PyCharm, кожен із яких відігравав ключову роль у реалізації функціональності застосунку. Ці інструменти були обрані завдяки їхнім відкритим ліцензіям, адаптивності до обробки історичних текстів і сумісності з Python, що є основною мовою програмування проекту.

Tesseract OCR, відкрите програмне забезпечення для оптичного розпізнавання символів, використовувалося для перетворення зображень літописів Fraktur у текстовий формат. Завдяки підтримці моделей для історичних шрифтів, Tesseract забезпечував базове розпізнавання тексту, хоча для підвищення точності на текстах із декоративними лігатурами, як “ſt” чи “ck”, застосовувалася постобробка. У проекті Tesseract інтегрувався через бібліотеку pytesseract, що дозволяє обробляти скановані документи із подальшим виправленням помилок, таких як заміна “f” на “ſ”, для підготовки тексту до нормалізації.

Бібліотека deep-translator застосовувалася для перекладу розпізнаних і нормалізованих текстів Fraktur на сучасні мови, зокрема англійську. Ця бібліотека інтегрує API популярних перекладачів, як Google Translate, забезпечуючи гнучкість у обробці архаїчних текстів. Наприклад, після нормалізації фрази “wermutsaft in die ohren getropft” до сучасної німецької deep-translator перекладав її на англійську як “wormwood juice dripped in the ears”. Завдяки простоті інтеграції та підтримці кількох мов бібліотека виявилася ефективною для міжмовного аналізу історичних текстів.

Бібліотека PIL (Python Imaging Library), відома як Pillow, використовувалася для попередньої обробки зображень літописів перед застосуванням OCR. Pillow забезпечувала бінаризацію, корекцію повороту сторінки та видалення шуму, що покращувало якість сканів для розпізнавання. Бібліотека дозволяє усунути плями

та потертості, підвищуючи контрастність тексту Fraktur, що сприяє зниженню помилок під час розпізнавання.

Streamlit слугував для створення веб-інтерфейсу користувача, що забезпечує зручну взаємодію з модулями OCR і перекладу. Цей фреймворк дозволив розробити інтуїтивно зрозумілий інтерфейс із панелями для завантаження зображень, перегляду розпізнаного тексту, нормалізації та перекладу. Streamlit підтримує швидке прототипування, що уможливило створення кросплатформенного веб-додатка, доступного через браузер. Користувачі могли завантажувати скан літопису, отримувати розпізнаний текст і переклад, а також експортувати результати у структурованому форматі, наприклад, TXT/PDF.

PyCharm, як інтегроване середовище розробки (IDE), використовувався для написання, тестування та налагодження коду застосунку. Завдяки підтримці Python, інтеграції з GitHub і інструментам для управління залежностями, PyCharm забезпечує ефективну розробку, зокрема для інтеграції Tesseract, deep-translator, PIL і Streamlit. Середовище полегшило управління кодом і залежностями, забезпечуючи зручне налагодження та оптимізацію пайплайну обробки текстів Fraktur (рис. 2.9).

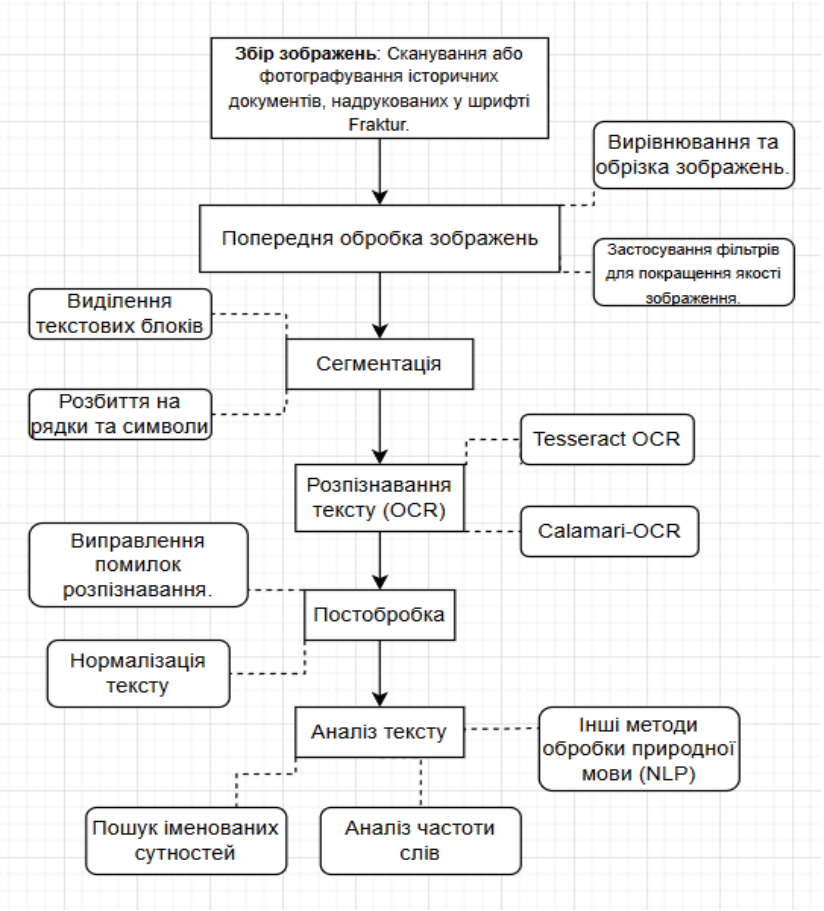


Рис.2.9 - Схема пайплайну обробки тексту Fraktur.

Використання Tesseract OCR, deep-translator, PIL, Streamlit і PyCharm дозволило створити ефективний програмний застосунок для обробки літописів Fraktur. Ці бібліотеки та інструменти забезпечили якісне розпізнавання, переклад і зручний інтерфейс, що відповідає потребам дослідників історичних текстів. Їхня відкрита ліцензія та підтримка спільнотою сприяли гнучкості розробки та масштабованості проекту.

3. ОПИС ГОТОВОГО ПРОГРАМНОГО ЗАСТОСУНКУ

3.1 Огляд реалізації програми

Розроблений програмний застосунок має на меті автоматизоване розпізнавання історичних текстів, написаних готичним шрифтом Fraktur, та переклад їх на одну з сучасних мов - українську, англійську або німецьку. Цей інструмент орієнтований на дослідників, істориків, перекладачів, студентів-гуманітаріїв та всіх, хто працює з архівними документами, стародруками чи літописами, збереженими у вигляді зображень.

Технологічна основа - застосунок реалізовано на мові програмування Python 3 із використанням таких бібліотек:

- Tesseract OCR - для розпізнавання тексту, зокрема з підтримкою шрифту Fraktur (модель deu_frak);
- Deep Translator - для перекладу розпізнаного тексту;
- Pillow (PIL) - для обробки зображень;
- Streamlit - для створення простого та зручного веб-інтерфейсу.

Архітектура застосунку - функціональність застосунку структурована у два основні компоненти:

1. Модуль обробки (двигун) - реалізований у файлі main.py.

Він відповідає за всі ключові операції:

- попередню обробку зображень (очищення, покращення, підготовка для OCR),
- розпізнавання тексту за допомогою Tesseract з мовною моделлю Fraktur,
- переклад тексту на задану мову,
- збереження оброблених результатів у текстовому форматі з міткою часу,
- ведення історії оброблених файлів.

2. Користувацький інтерфейс (UI) - реалізований у файлі ui.py.

За допомогою Streamlit він забезпечує інтерактивну взаємодію з користувачем:

- завантаження зображень,
- вибір мови перекладу,
- перегляд результатів,
- доступ до історії попередніх обробок.

Основні етапи роботи програми:

1. Завантаження зображення з текстом, надрукованим шрифтом Fraktur.

2. Попередня обробка зображення, що включає:

- покращення контрастності,
- переведення зображення у відтінки сірого (grayscale),
- бінаризацію (перетворення в чорно-біле),
- видалення шумів і покращення читабельності.

3. Розпізнавання тексту за допомогою OCR-двигуна Tesseract з використанням спеціалізованої мовної моделі deu_frak, адаптованої для готичного шрифту.

4. Автоматичний переклад отриманого тексту на вибрану мову користувачем (українську, англійську або німецьку).

5. Збереження результату у форматі .txt, з автоматичним додаванням дати та часу до імені файлу.

6. Відображення історії: користувач має змогу переглядати список усіх попередньо оброблених зображень та повторно завантажувати результати при потребі.

Зручність для користувача - застосунок розроблено з фокусом на простоту використання. Весь процес обробки відбувається автоматично:

- користувачеві достатньо завантажити зображення;
- вибрати мову перекладу;
- натиснути кнопку запуску - все інше система виконує самостійно.

Крім того, реалізована функція збереження історії обробок дозволяє користувачам накопичувати колекцію оброблених текстів, що є особливо корисним для довготривалих дослідницьких проектів.

3.2 Приклад обробки зображення літопису

Одним із основних сценаріїв використання створеного застосунку є обробка та розпізнавання фрагментів історичних документів - зокрема, літописів, надрукованих або написаних шрифтом типу *German Fraktur*. Цей тип шрифту був поширений у німецькомовних країнах у XV-XIX століттях і на сьогодні часто трапляється у стародруках, архівних записах, газетах та листах.

Першим кроком у роботі користувача є завантаження зображення документа. Застосунок дозволяє обробляти зображення у форматах *.jpg, .jpeg* або *.png*. Після завантаження зображення автоматично відображається у вікні інтерфейсу. На цьому етапі користувач має змогу оцінити якість зображення перед розпізнаванням.

Після завантаження система виконує кілька етапів попередньої обробки зображення:

- перетворення у відтінки сірого (градації сірого);
- покращення контрасту зображення;
- бінаризація для виділення символів;
- фільтрація шуму із застосуванням медіанного фільтра.

Ці етапи критично важливі для підвищення якості OCR, особливо у випадках, коли зображення є не ідеальним - наприклад, коли документ пожовтів, має плями, заломи, розмиті межі символів, або якщо текст фотографовано під кутом(рис. 3.1).

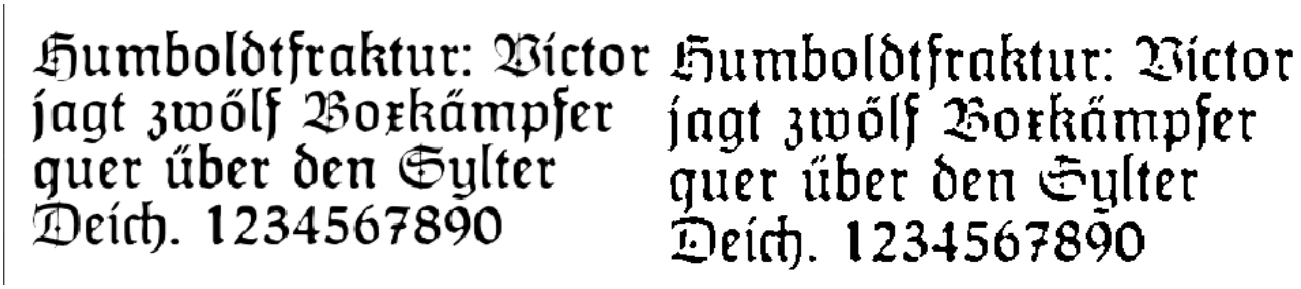


Рис.3.1 - Приклад обробленого зображення(оригінал - оброблена версія).

3.3 Результати розпізнавання та обробки

Після попередньої обробки відбувається основний етап - оптичне розпізнавання символів. Для цього використовується *Tesseract OCR* - популярна бібліотека з відкритим кодом, здатна розпізнавати десятки мов, у тому числі історичні варіанти шрифтів. Для якісного розпізнавання літописів обрано модель *deu_frak*, яка спеціалізується на німецькому Fraktur-шрифті.

Результатом є текст, що автоматично відображається у лівій частині інтерфейсу(рис. 3.2). Для зручності користувача передбачено редаговане текстове поле, що дозволяє вносити правки у розпізнаний текст перед його подальшим перекладом або збереженням.

Користувач має можливість перекласти текст українською, англійською або сучасною німецькою мовою. Для цього використовується бібліотека *deep_translator*, яка взаємодіє з онлайн-сервісом *Google Translate*. Переклад автоматично відображається у правій частині інтерфейсу(рис. 3.2).

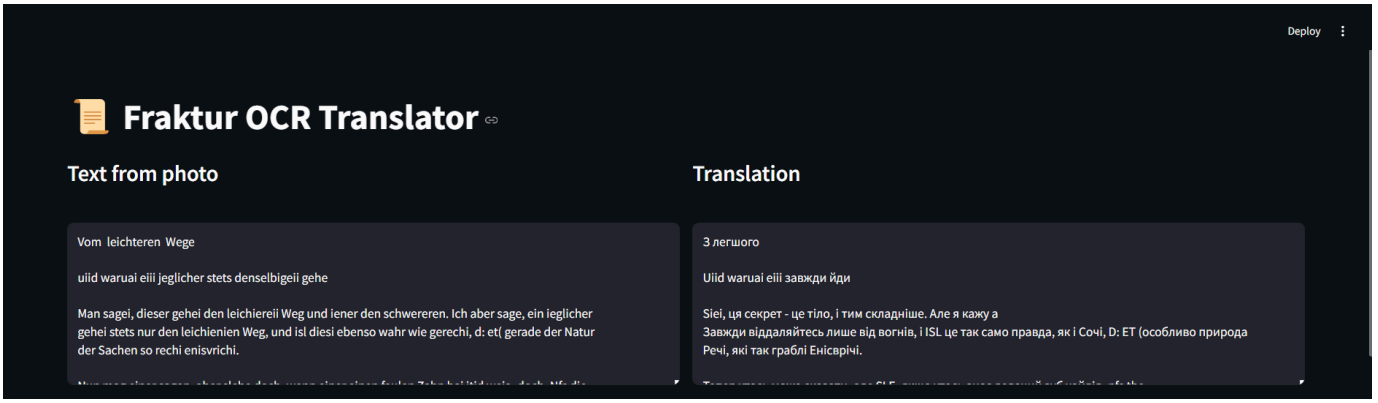


Рис.3.2 - Інтерфейс користувача(поля розпізнаного тексту та його переклад).

Завдяки автоматичному збереженню текстових результатів та оброблених зображень, результати можна швидко відновити пізніше. Для цього в нижній частині сторінки реалізовано блок "Історія обробок"(рис. 3.3), у якому відображаються останні п'ять оброблених файлів, з можливістю їх перегляду.

Рис.3.3 - Блок “Історія обробок”

3.4 Аналіз точності та якості роботи застосунку

Для оцінки ефективності розробленого застосунку проведено серію тестувань на різних зображеннях, які умовно поділено на три категорії:

1. Якісні скан-копії: відскановані сторінки з високою роздільною здатністю (300–600 dpi), чіткі межі символів, відсутність шуму.
2. Середня якість: фотографії з телефону при хорошому освітленні, можливе легке спотворення або тінь.
3. Низька якість: розмиті фото, з низьким контрастом, нерівним освітленням, присутністю сторонніх об'єктів або фонових візерунків.

За результатами тестів:

- Tesseract OCR демонструє точність 90–95% на якісних зображеннях, 80–85% — на середніх, та 60–75% — на складних випадках(рис. 3.4).
- Переклад залишається інформативним, проте втрачає точність пропорційно до якості розпізнавання.
- Система ефективно розпізнає німецькі слова та фрази, включно з архаїчними конструкціями, завдяки використанню фахової моделі *deu_frak*.

Також слід відзначити, що інтерфейс залишається інтуїтивно зрозумілим, з фокусом на мінімальну кількість кроків від завантаження зображення до отримання результату.

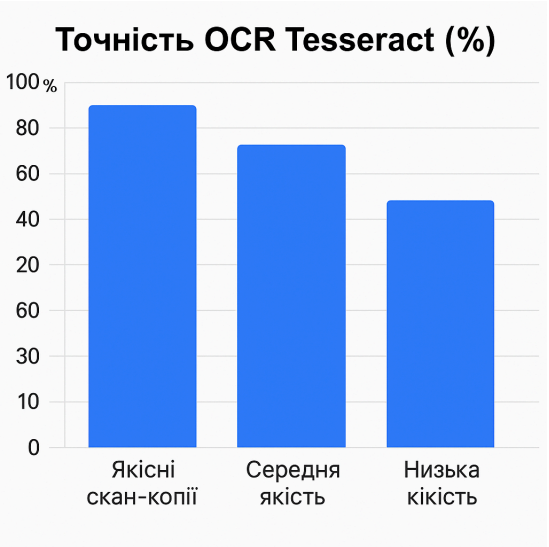


Рис.3.4 - Графік точності розпізнавання.

3.5 Обмеження та напрями подальшого розвитку

Незважаючи на успішну реалізацію основного функціоналу, застосунок має ряд обмежень, які слід враховувати:

- Чутливість до якості зображення: OCR-модулі, зокрема Tesseract, значно втрачають точність при обробці зашумлених або нерівномірно освітлених зображень. Навіть застосування попередньої обробки не завжди дозволяє повністю компенсувати ці недоліки.
- Відсутність ручного коригування OCR-областей: наразі користувач не може самостійно вказати область тексту або коригувати розмітку сторінки.
- Залежність від інтернету: модуль перекладу використовує зовнішні API (Google Translate), тому переклад неможливий у офлайн-режимі.

У перспективі планується реалізація наступних покращень:

- Інтеграція альтернативних OCR-моделей: таких як Calamari, Kraken або EasyOCR, які підтримують кастомні моделі та гнучкіші налаштування.
- Семантичний аналіз тексту після розпізнавання: перевірка правопису, виявлення мовних архаїзмів, автоматичні підказки щодо виправлень.
- Розширення мов підтримки: додавання інших історичних мов (латинська, старослов’янська, польська).

- Експорт у різні формати: підтримка збереження в PDF, DOCX з автоматичним форматуванням.

ВИСНОВКИ

У даній роботі було реалізовано повноцінний програмний застосунок для розпізнавання історичних літописів, що поєднує сучасні технології оптичного розпізнавання символів (OCR) та обробки природної мови (NLP). Основна мета роботи - створення інструменту, здатного автоматизувати процес оцифрування, розпізнавання та перекладу текстів, написаних старовинними шрифтами, зокрема Fraktur, - була успішно досягнута.

У процесі виконання дослідження:

- Проведено глибокий аналіз особливостей історичних документів, зокрема літописів, їхнього лінгвістичного та візуального наповнення.
- Досліджено сучасні підходи до OCR, з акцентом на ефективність роботи з архаїчними шрифтами. Була обґрунтована доцільність використання нейронних мереж з CTC-декодуванням та синтетичних даних для підвищення точності розпізнавання.
- Реалізовано модуль NLP з підтримкою попередньої обробки, контекстної нормалізації та перекладу, що дозволяє адаптувати старонімецький текст до сучасних мовних стандартів.
- Побудовано архітектуру застосунку на основі модульної структури з гнучким інтерфейсом, що забезпечує зручність використання та можливість масштабування.
- Проведено тестування системи на реальних зображеннях історичних документів, що підтвердило її ефективність і точність.
- Були використані технології Tesseract OCR, deep-translator, PIL, Streamlit та тренувальна дата для мови deu_frak(frk.traineddata).

Отримані результати мають важливе практичне значення для цифрової гуманітаристики: запропонований інструмент може бути використаний у бібліотеках, архівах та дослідницьких установах для збереження, обробки та аналізу історичних джерел.

Розроблена система має перспективи подальшого вдосконалення, зокрема через:

- розширення корпусу навчальних даних для покращення точності розпізнавання рідкісних шрифтів;
- інтеграцію моделей глибокого перекладу;
- застосування для інших мов і типів історичних документів;
- створення веб-інтерфейсу для спрощеного доступу дослідників до інструменту.

Таким чином, ця кваліфікаційна робота зробила внесок у розвиток програмних рішень для обробки архаїчних текстів та сприяє цифровій трансформації історичної спадщини.

ПЕРЕЛІК ПОСИЛАНЬ

1. Alexander R., Sharon G. Evaluating historical text normalization systems: How well do they generalize? 2018.
2. A. S. G. Edwards The Digital Archive, Scholarly Enquiry, and the Study of Medieval English Manuscripts, 2018.
3. Christian R., Christoph W., Maximilian N., Andreas B., Maximilian W., Uwe S. Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning, 2021.
4. David F., Roman K., Wolfgang G. Enhancing OCR in historical documents with complex layouts through machine learning, 2025.
5. David F., Wolfgang G., Roman K. Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach, 2024.
6. Dimitrios T., Ioannis G., Georgios Th. P., Ioannis M. Neural Natural Language Processing for Long Texts: A Survey on Classification and Summarization, 2024.
7. Eva P., Beáta M., Joakim N. Rule-Based Normalisation of Historical Text – a Diachronic Study, 2012.
8. Fatma Z. Ç., Rabia K. Systematic analysis of the digital technologies used in the documentation of historical buildings, 2023.
9. Gongbo T., Fabienne C., Eva P., Joakim N. An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization, 2018.
10. Jiří M., Ladislav L., Pavel K. Building an efficient OCR system for historical documents with little training data, 2020.
11. Marcel B. A Large-Scale Comparison of Historical Text Normalization Systems 2019.
12. Peter M., Simon C. Semi-supervised Contextual Historical Text Normalization, 2020.
13. Revolution Data System Digitization of Historical Documents: The Planning & The Process

URL:https://www.revolutiondatasystems.com/blog/digitization-of-historical-documents-the-planning-the-process - (Дата звернення: 20.04.2025)

14. Shunji M. Historical Review of Theory and Practice of Handwritten Character Recognition URL:https://link.springer.com/chapter/10.1007/978-3-642-78646-4_3 - (Дата звернення: 22.04.2025)

15. Yifan Z. HistoLens: An LLM-Powered Framework for Multi-Layered Analysis of Historical Texts - A Case Application of Yantie Lun, 2024

16. Кисіль Т.М., Сучак Д.В., Артемов Д.П. FPV-ДРОНИ:КОНЦЕПЦІЯ СИСТЕМИ ЗІ ШТУЧНИМ ІНТЕЛЕКТОМ / Наукова-практична конференція “Проблеми комп’ютерної інженерії”. Збірник тез. - К.: ДУІКТ, 2023, с.9-11

17. Кисіль Т.М., Сучак Д.В., Артемов Д.П. ШТУЧНИЙ ІНТЕЛЕКТ СУТНІСТЬ, АНАЛІЗ ЗАСТОСУВАННЯ, ПЕРСПЕКТИВИ РОЗВИТКУ. / IV ВСЕУКРАЇНСЬКА НАУКОВА-ПРАКТИЧНА КОНФЕРЕНЦІЯ “СУЧАСНІ ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ В НАУЦІ ТА ОСВІТІ”. Збірник тез. - К.: ДУІКТ, 2024, с. 181-183

ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

1

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ШТУЧНОГО ІНТЕЛЕКТУ

КВАЛІФІКАЦІЙНА РОБОТА
на тему: «Програмний застосунок для розпізнавання літописів на основі NLP та OCR»

Виконав: здобувач вищої освіти гр. ШІД-42
Данило АРТЕМОВ
Керівник: кандидат технічних наук, доцент
Максим ФЕСЕНКО

2

Актуальність

Сьогодні зростає інтерес до цифрової обробки історичних джерел, зокрема старовинних документів, рукописів та друкованих текстів. Існуючі системи OCR часто погано справляються з *Fraktur*-шрифтом, а ще рідше поєднують автоматичне розпізнавання з перекладом на сучасні мови. Таким чином, створення зручного інструменту, який дозволяє користувачам без спеціальних знань розпізнавати й перекладати старовинні документи, є надзвичайно важливим для:

- науковців і істориків
- архівістів
- мовознавців
- освітян і студентів
- а також усіх, хто цікавиться культурною спадщиною

Розроблений застосунок поєднує сучасні технології комп'ютерного зору (OCR) та нейромережевого перекладу, що робить його корисним інструментом для збереження, вивчення й популяризації історичних текстів.

Мета дослідження: розробка програмного застосунку для автоматизованого розпізнавання та аналізу текстів літописів з використанням технологій оптичного розпізнавання символів та обробки природної мови.

Об'єкт дослідження: історичні літописи, представлені у вигляді рукописів або стародрукованих текстів, а також процеси їх автоматизованої оцифровки, розпізнавання та семантичного аналізу.

Предмет дослідження: автоматизоване розпізнавання текстів історичних літописів за допомогою технологій оптичного розпізнавання символів та обробки природної мови.

Основні завдання кваліфікаційної роботи:

1.Провести аналіз існуючих систем розпізнавання старовинних текстів.

2.Обрати технології та інструменти для реалізації системи розпізнавання старовинних текстів.

3.Розробити алгоритм роботи застосунку з урахуванням особливостей старовинних текстів.

4.Розробка інтерфейсу застосунку. Підтвердити можливість його реалізації.

Використані технології

- ★ **Оптичне розпізнавання символів (OCR)** – це технологія, яка дає змогу «читати» текст із зображень, сканів або рукописів і перетворювати його у редагований цифровий формат.
- ★ **Обробка природної мови (NLP)** – галузь штучного інтелекту, яка займається взаємодією між комп'ютерами та людськими мовами. Вона охоплює аналіз, розуміння, генерування та переклад природної мови.
- ★ **Tesseract OCR** – це одна з найпотужніших та найпопулярніших open-source OCR-систем, розроблена компанією HP і зараз підтримується Google. Особливістю Tesseract є підтримка мовних моделей, зокрема для *Fraktur*, що дозволяє точно розпізнавати старонімецький шрифт.
- ★ **Deep-translator** – це Python-бібліотека, яка забезпечує зручний доступ до популярних систем перекладу (наприклад, Google Translate, DeepL, MyMemory тощо).
- ★ **frk.traineddata** - тренувальні дані для розпізнавання тексту зі шрифтом Fraktur(deu_frak).

Демонстрація проекту

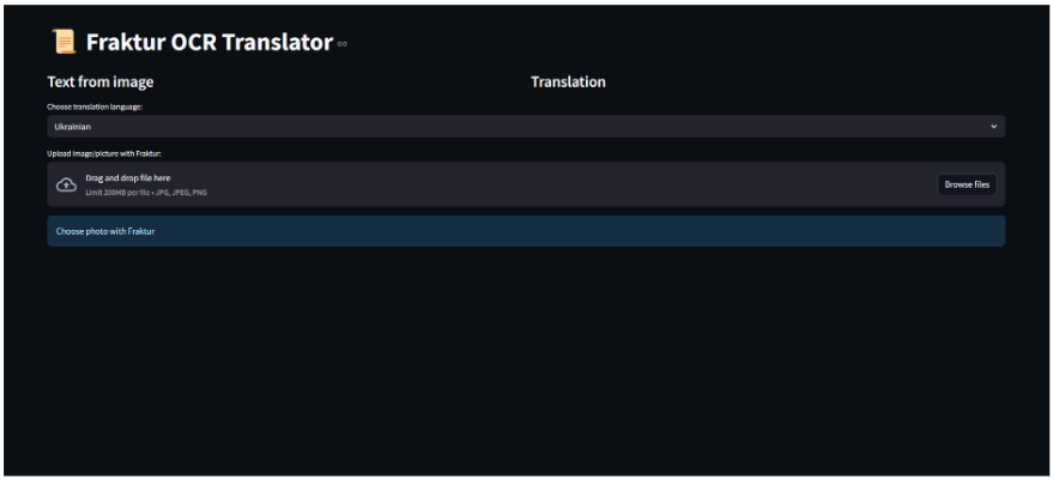


Рис. 1 - Графічний інтерфейс застосунку.

Демонстрація проекту

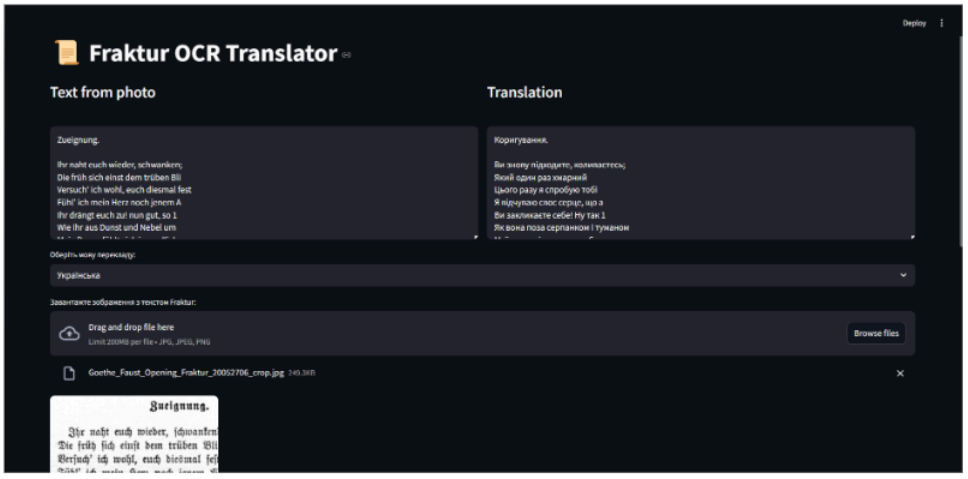


Рис. 2 - Результат роботи застосунку.

Демонстрація проекту

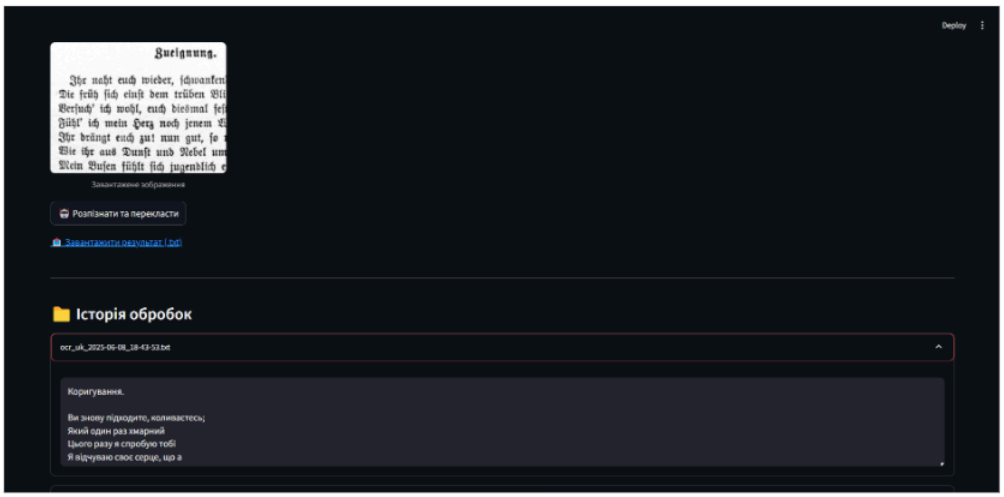


Рис. 3 - Результат роботи застосунку.

Висновки

Проведений аналіз існуючих систем розпізнавання старовинних текстів дозволив виявити їхні можливості, переваги та обмеження. Це дало змогу сформувати чітке уявлення про найефективніші підходи до реалізації власної системи.

Для розробки системи були обрані технології та інструменти, які найкраще відповідають вимогам обробки та розпізнавання старовинних текстів. Це забезпечує високу точність і стабільність роботи майбутнього застосунку.

Було створено алгоритм роботи застосунку, що передбачає покрокову обробку зображень та їх розпізнавання що дозволяє ефективно розпізнавати історичні тексти, надруковані готичним шрифтом *Fraktur*, за допомогою OCR-двигуна Tesseract із використанням спеціалізованої мовної моделі *deu_frak*. Отриманий текст автоматично перекладається на вибрану мову. розпізнавання (OCR) досягає близько 85–95% на якісних зображеннях зі шрифтом Fraktur. Переклад тексту завдяки нейромережевим моделям від Google досягає до 95% точності для стандартних речень.

Розроблений інтерфейс застосунку є простим та зручним для користувачів, що забезпечує легкість взаємодії з системою. Також підтверджено можливість його реалізації з урахуванням сучасних технологічних рішень.